

Computer biology system

Also published as:

 DE19642651 (B4)

US5787279 (A)

US5787279 (A)

US5787279 (A)

US5787279 (A)

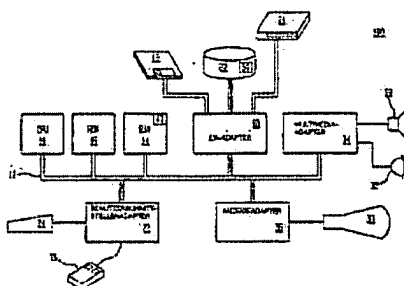
US5787279 (A)

US5787279 (A)

US5787279 (A)

Abstract of DE 19642651 (A1)

To store the presentation of N reference molecule in a memory of a computer system, N rigid substructure of the reference molecule is detected, where each substructure has N atom positions. Each atom position is linked with N atom position in the rigid substructure by a fixed link. Each rigid substructure has a global position and a global orientation in a global coordinate system. At least 2 vectors are defined with a value and direction with a fixed position and orientation in relation to a selected rigid substructure from the rigid substructures. A set of N positions is selected in the selected rigid substructure to form positions of a multiple coordinate system where N of the positions is not co-linear with the remaining positions, and the positions are fixed in relation to the selected rigid substructure. The multiple coordinate system defines a three-dimensional and angled local coordinate system. One or more of the multiple coordinate systems are selected to develop a multiple coordinate system field with data which can be coupled to each selected multiple coordinate system. A data set is stored in a data structure containing a number of data sets each containing a multiple coordinate system field and a vector field. The vector field has vector data which can be related to each of the vectors together with data on the molecule identities and the selected rigid substructure. Also claimed is a computer system to display N reference molecule with a memory and the facility to compare N reference molecule with a test molecule. A database is stored in memory with a display of N rigid substructures of each reference molecule.



Data supplied from the *esp@cenet* database — Worldwide

19 BUNDESREPUBLIK
DEUTSCHLAND



DEUTSCHES
PATENTAMT

12 Offenlegungsschrift
10 DE 196 42 651 A 1

51 Int. Cl. 8:
G 06 F 19/00
G 06 F 17/30
A 61 K 31/00
// G 06 F 159:00

21 Aktenzeichen: 196 42 651.0
22 Anmeldetag: 16. 10. 96
43 Offenlegungstag: 28. 6. 97

DE 196 42 651 A 1

31 Unionspriorität: 32 33 31
22.12.95 US 577353

71 Anmelder:
International Business Machines Corp., Armonk,
N.Y., US

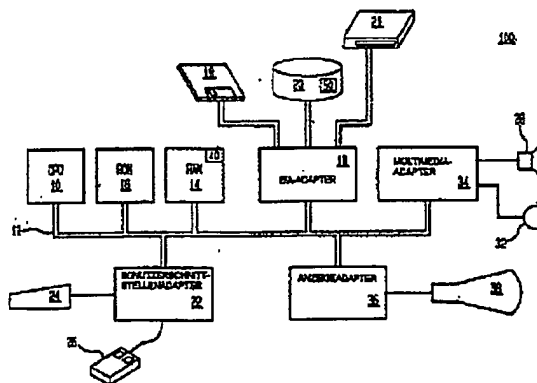
74 Vertreter:
Schäfer, W., Dipl.-Ing., Pat.-Anw., 70188 Stuttgart

22 Erfinder:
Rigoutsos, Isidore, Long Island City, N.Y., US

Prüfungsantrag gem. § 44 PatG ist gestellt

54 System und Verfahren zur Erkennung konformationsflexibler Moleküle

51 Ein Referenzspeicherprozeß bevölkert eine Datenstruktur, derart, daß die Datenstruktur alle molekularen Strukturen und/oder starren Substrukturen in der Datenbank enthält, die gemäß Eigenschaften von Tupeln klassifiziert sind. Bei einer bevorzugten Ausführungsform werden die Tupel von Plätzen (z. B. Atomplätzen) der molekularen Strukturen abgeleitet, und die Eigenschaften können von geometrischen (und anderen) Informationen abgeleitet werden, die mit den Tupeln in Beziehung stehen. Die Eigenschaften werden dazu verwendet, Indizes in der Datenstruktur zu definieren, die mit invarianter Vektorinformation (z. B. Information über drehbare Bindung(en) in schiefwinkligen lokalen Koordinatensystemen, die von Tupeln erzeugt werden) verknüpft sind. Diese Darstellungen sind invariant bezüglich der Rotation und Translation von molekularen Strukturen und/oder der Rotation von Substrukturen um angebundene drehbare Bindungen. Demgemäß wird die invariante Vektorinformation in der Datenstruktur klassifiziert, wobei sich die jeweiligen Tupel Eigenschaften an Stellen befinden, die durch den von dem jeweiligen Tupel abgeleiteten Index bestimmt werden. Ein Vergleichsprozeß erzeugt ein oder mehr Tupel, schiefwinklige lokale Referenzkoordinatensysteme und Indizes (Testkoordinatensystemtupelindizes genannt) für die Struktur (Substrukturen) eines Testmoleküls unter Verwendung der gleichen Technik, die zum Bevölkern der Datenstruktur verwendet wurde. Der Testkoordinatensystemtupelindex greift ...



DE 196 42 651 A 1

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen

BUNDESDRUCKEREI 04. 97 702 026/651

49/24

GEBIET DER ERFINDUNG

5 Diese Erfindung bezieht sich auf das Gebiet der Rechnerbiologie. Spezieller bezieht sich die Erfindung auf ein Rechnersystem und ein Verfahren zur Erkennung jener Moleküle in einer Datenbank aus einem oder mehreren Molekülen, die gemeinsam mit einem oder mehreren Testmolekülen Substrukturen enthalten, auch wenn die Moleküle in der Datenbank Gruppen von Atomen enthalten, die sich frei um jegliche kovalente Bindungen
10 herum drehen können, die möglicherweise in derartigen Molekülen existieren (Torsionsflexibilität).

HINTERGRUND DER ERFINDUNG

Da vorhandene Informationsdepots schneller bearbeitet werden müssen und eine größere Vielzahl an Werkzeugen verfügbar wird, spielt der Rechner eine zunehmend wichtigere Rolle bei der Führung und Rationalisierung
15 des Arzneimittellentdeckungs- und -entwurfsprozesses.

Einer der grundlegenden Bestandteile jüngerer Vorgehensweisen auf dieser Linie von Forschungsbemühungen war der Wunsch, molekulare Eigenschaften, die auf den grundlegendsten Niveaus von Arzneimittelwechselwirkung involviert sind, zu berechnen, zu katalogisieren und nach ihnen zu suchen.

20 Speziell können Rechner den Forschern helfen, rasch a priori aussichtslose Kandidaten zu eliminieren, um so langwierige und kostenintensive Aktivitätssichtungen zu vermeiden. Wichtiger, sie können es Forschern ermöglichen, neue vielversprechende Verbindungen zu identifizieren, lediglich basierend auf der verfügbaren Information auf der Rezeptorstelle oder auf anderen Führungsverbindungen.

Indem man in der Lage ist, diese Aufgaben schnell durchzuführen und Information zu gewinnen, die sofort in
25 das Ausgangsgemisch des Arzneimittels eingebracht werden kann, wird erwartet, daß die Suchstrategie dieses komplexe, multidisziplinäre Bemühen stark vereinfacht und die Geschwindigkeit, mit der neue und effektivere Arzneimittel identifiziert, getestet und auf den Markt gebracht werden, signifikant erhöht.

Bis heute wurden Hunderte von Proteinstrukturen mittels Röntgenstrahlkristallographie- und magnetischen Kernresonanz-(NMR)-Verfahren bestimmt. Diese Daten sind ohne weiteres als eine öffentliche Ressource von
30 Molekularstrukturdaten verfügbar und erlauben es Pharmakologen und Biologen, verschiedene Aspekte von Proteinstrukturen und deren komplexe Verhaltensweisen zu untersuchen. Zusätzlich zu diesen öffentlichen Datenbanken wurde eine Anzahl von weiteren (öffentlichen und privaten) Datenbanken von kleinen organischen Molekülen durch die Anstrengungen von zahlreichen pharmazeutischen und biotechnologischen Firmen und Forschungsorganisationen aufgebaut.

35 Es gibt verschiedene unterschiedliche Szenarien, auf die man wahrscheinlich im Prozeß des Entwurfs von Arzneimitteln stößt:

1. Es wird ein pharmakophores Modell von verschiedenen aktiven Molekülen vorgeschlagen; man wünscht, andere Moleküle zu finden, welche die pharmakophore Hypothese entweder unterstützen oder widerlegen.
- 40 2. Eine Anzahl ungeprüfter Moleküle kann biologische Aktivität zeigen; man wünscht, vorhandene Beziehungen zwischen dreidimensionaler Struktur und Aktivität auszuwerten, um potentiell vorhandene biologische Eigenschaften abzuleiten.
3. Es wurde vorgeschlagen, daß eine bestimmte Konformation eines gegebenen Liganden biologisch aktiv ist; es wird angenommen, daß eine dreidimensionale Suche weitere Moleküle identifiziert, die den Liganden
45 nachahmen.
4. Die dreidimensionale Struktur einer Protein- oder DNA-Bindungsstelle steht durch kristallographische Untersuchungen zur Verfügung. In diesem Fall werden Liganden gesucht, die zu der Bindungsstelle passen. Die Suche kann durch Verwenden von Information über das elektrostatische Verhalten für einige der Atomgruppen an der Stelle vergrößert werden. Diese Technik ist von mehreren pharmazeutischen Firmen,
50 wie Merck, American Cyanamid, Agouron etc. erfolgreich angewendet worden, um Liganden zu entwerfen, die das reverse Transkriptase-Enzym von HIV binden und blockieren.

In jedem dieser Fälle muß die entscheidende Information, welche Forscher in die Lage versetzt, Hypothesen zu entwickeln, die potentiell neue Molekülkandidaten für Synthese und Prüfung betreffen, durch eine Suche in
55 einer potentiell sehr großen Datenbank relevanter Information gewonnen werden. Tatsächlich ist das mehreren Stufen medizinischer Chemieuntersuchungen gemeinsam zugrundeliegende Element die Notwendigkeit eines Suchens nach Datenbanken für chemische Information. Das Folgende konzentriert sich auf den Fall, in dem die zu suchenden Datenbanken strukturelle Information enthalten, die zu 3-dimensionalen atomaren Verbindungen gehören.

60 Typischerweise liegt einem eine Verbindung/ein Molekül C in Form eines Satzes von Koordinaten der Atomplätze der Verbindung vor. Außerdem ist eine Datenbank D, d. h. eine Sammlung von Sätzen $D_j = \{ \dots \}$, gegeben. $D_j = \{ \dots \}$ ist eine Sammlung von Koordinatensätzen der Atomplätze für jedes der beteiligten Moleküle. Bindungen, von denen einige drehbar sein können und somit für eine Torsionsflexibilität sorgen, verbinden die verschiedenen Atomplätze sowohl in C als auch den Datenbankmitgliedern. Torsionsflexibilität bedeutet, daß
65 die Gruppen von Atomen, die an den zwei Endpunkten einer (drehbaren) Bindung starr angebunden sind, relativ zueinander rotieren können. Jede Verbindung/jedes Molekül kann mehr als eine drehbare Bindung enthalten, und somit kann die Verbindung/das Molekül eine beliebige einer unbegrenzten Anzahl von Konformationen (dreidimensionalen Konfigurationen) über Drehungen um diese Bindungen herum annehmen. Gelegentlich

können sterische Randbedingungen oder Energiebetrachtungen die Anzahl an Wahlmöglichkeiten begrenzen, die Hauptmenge des Satzes an möglichen Konfigurationen bleibt jedoch nichtsdestoweniger unbegrenzt. Diese Konformationsflexibilität molekularer Strukturen eröffnet einen breiten Bereich von Möglichkeiten bei der Suche nach potentiellen Liganden, während sie gleichzeitig das Problem exponentiell schwieriger macht. Zusätzlich zu der (internen) Torsionsflexibilität ist es den Molekülen möglich, starre Transformationen im dreidimensionalen Raum zu erfahren, d. h. das Molekül als ganzes kann sich drehen und verschieben. Im folgenden wird die Verbindung/das Molekül C austauschbar als 'Testverbindung' oder 'Testmolekül' oder 'Abfrageverbindung' oder 'Abfragemolekül' bezeichnet.

Bei einer gegebenen Verbindung C und einer gegebenen Datenbank D, die Informationen über die 3-dimensionale Struktur eines möglicherweise großen Molekülsatzes enthält, müssen die folgenden Operationen definiert und ausgeführt werden:

1. "Struktureinfügung": die Fähigkeit, sämtliches zur Verfügung stehendes strukturelles Wissen über die Verbindung C in die Datenbank D einzubauen;
2. "Strukturzugehörigkeit": Bestimmung, ob die Verbindung C bereits in der Datenbank D enthalten ist;
3. "Substruktursuche": Identifizieren und Berichten aller Mitgliedsverbindungen von D, die eine bestimmte Substruktur der Verbindung C enthalten;
4. "Ähnlichkeitssuche": Identifizieren und Berichten aller Mitgliedsverbindungen von D, die der Verbindung C ähnlich sind. Um eine derartige Operation in der Datenbank D zu implementieren, muß ein Ähnlichkeitsmaß $d(\dots)$ definiert werden und zur Verfügung stehen, und
5. "Überstruktursuche": Identifizieren und Berichten aller Mitgliedsverbindungen von D, die eine Substruktur der Verbindung C darstellen.

Als erstes ist es leicht einzusehen, daß das Prädikat der Strukturzugehörigkeit der Substruktursuchoperation untergeordnet ist. Des weiteren können alle Suchoperationen auf das reduziert werden, was wir als "Substrukturähnlichkeit" bezeichnen wollen.

Im folgenden wird der Ausdruck Substrukturähnlichkeit dazu verwendet, eine einzelne Operation zu bezeichnen, die, wenn eine Verbindung C, eine Datenbank D und ein Ähnlichkeitsmaß $d(\dots)$ gegeben sind, die Bestimmung aller Verbindungsmitglieder von D erlaubt, die eine Substruktur enthalten, die einer Substruktur von C ähnlich ist. Das Ausmaß an Ähnlichkeit zwischen den fraglichen Molekülen kann durch die Funktion $d(\dots)$ bestimmt werden. Dies ist hier so zu verstehen, daß die implizierte gemeinsame Substruktur nicht notwendigerweise ein eigentlicher Subsatz von C sein muß. Die Ähnlichkeitsfunktion $d(\dots)$ bleibt unspezifiziert, wir nehmen jedoch an, daß sie von einer sehr allgemeinen Natur ist.

Von dem Problem des Substrukturvergleichs kann gezeigt werden, daß es NP-vollständig ist, indem bemerkt wird, daß es das Problem des Subgraph-Isomorphismus als einen Spezialfall beinhaltet. Die reale Folge dieser Aussage besteht darin, daß die Zeitkomplexität zum Auffinden aller optimalen Lösungen eine exponentielle Funktion der Eingabelänge ist, und somit existiert kein effizienter Algorithmus (d. h. mit polynomialer Zeitkomplexität) zum Auffinden optimaler Lösungen. Zur Berechnungskomplexität des Problems trägt des weiteren bei, daß eine Torsionsflexibilität um die kovalenten Bindungen des Moleküls herum ermöglicht wird.

Vor der Beendigung dieses Abschnitts sollte eine letzte Unterscheidung erwähnt werden. Dies ist die Unterscheidung zwischen 'Identifizierung' und 'Erkennung' jener Moleküle von der Datenbank D und Molekülen, die der gegebenen Testverbindung/dem gegebenen Testmolekül C ähnlich sind. Die Identifizierung beschränkt sich selbst darauf, lediglich die Identitäten der Moleküle von der Datenbank D zu berichten, die mit der Testverbindung/dem Testmolekül C zusammenpassen. Andererseits erfordert eine Erkennung nicht nur das Berichten der Identitäten der zusammenpassenden Moleküle, sondern auch die Bestimmung und das Berichten der notwendigen Transformationen, die jedes der identifizierten zusammenpassenden Moleküle in "beste Übereinstimmung" mit der Testverbindung/dem Testmolekül bringen. ("Beste Übereinstimmung" positioniert die Atomplätze des Testmoleküls und der dazu passenden Datenbankmoleküle in einer solchen Weise, daß die Anzahl von Plätzen im dreidimensionalen Raum, die gleichzeitig von Atomen des Testmoleküls und Atomen jedes dazu passenden Datenbankmoleküls besetzt werden, die maximal mögliche ist.) Diese "notwendigen" Transformationen umfassen Rotationen und Translationen der betrachteten Moleküle als ganzes, jedoch auch Rotationen von Strukturen innerhalb der Moleküle um die torsionsflexiblen Bindungen der Moleküle herum.

Erkennung ist von der Sache her ein viel schwierigeres Problem als Identifizierung; dies ist im Fall von sehr großen Datenbanken D mit Molekülen, die torsionsflexibel sind, besonders offensichtlich. Dies ist so, weil die Anzahl von möglichen Transformationen mit der Anzahl von drehbaren Bindungen, die für die gezeigte Torsionsflexibilität sorgen, exponentiell zunimmt: die Berechnungsvorgänge zum Finden und Berichten der richtigen Transformation nehmen im allgemeinen mit der Anzahl von Transformationen zu.

PROBLEME BEIM STAND DER TECHNIK

Die inhärente Berechnungskomplexität der Fragestellung bezüglich Substrukturähnlichkeit hat typisch alle zuvor vorgeschlagenen Vorgehensweisen zur Bewältigung dieses Problems belastet. Auch wenn man das Problem auf den Fall von starren Molekülen ohne rotierbare Bindungen beschränkt, bleibt das Problem aufgrund seiner dreidimensionalen Natur mit einer hohen Anforderung für die Berechnung.

Um die Komplexität der Fragestellung richtig einzuschätzen, wird ein eindimensionales Analogon aus dem täglichen Leben präsentiert. Ist ein Regal voll von Büchern und ein Satz wie

"Bilden von 3D-Anfragen, die sich auf eine bestimmte Flexibilität in den Targetstrukturen einstellen können,"

gegeben, so ist eine Suchaufgabe als die Notwendigkeit definiert, jedes Auftreten ähnlicher Redewendungen in dem Satz verfügbarer Bücher zu finden. Ähnlich bedeutet, daß im allgemeinsten Fall eine Redewendung wie

"wir bilden eine 3D-Suchanfrage derart, daß sie sich auf die gewünschte Flexibilität einstellt"

als ein gültiges Zusammenpassen berichtet werden sollte. Mit anderen Worten sind Operationen wie Ersetzung, Einfügung und Löschung der grundlegendsten Informationselemente (in diesem Fall der Buchstaben) gerechtfertigt und somit erlaubt. (Im Fall von Molekülen sind die grundlegendsten Informationselemente die Atome eines Moleküls.)

Eine direkte Vorgehensweise zur Lösung dieser Aufgabe erfordert die Erfassung der Inhalte aller Bücher des Regals in einer erschöpfenden, linearen Weise, d. h. von links nach rechts, von oben nach unten, um alle ähnlichen (in diesem Fall eindimensionalen) Strukturen zu lokalisieren. Klarerweise erfordert ein derartiger Operationsmodus mit zunehmender Anzahl von Büchern auf dem Regal (d. h. Größe der Datenbank) zunehmend mehr Zeit.

Selbstverständlich kann es eine Anzahl heuristischer, Betrachtungen erleichtern, die Antwort zu finden. Zum Beispiel sind möglicherweise bestimmte Operationen nicht erlaubt, oder die Suche kann auf einen kleineren, gut spezifizierten Satz beschränkt werden; dies begrenzt unmittelbar die Anzahl möglicher Varianten für eine gegebene Redewendung und macht eine Vorausberechnung und ein Speichern alternativer Redewendungen plausibel. Beim Suchen wird die Testredewendung mit dem Satz aller erlaubten, vorausberechneten Varianten verglichen.

Alternativ können "Schlüssel" unter Verwendung eines Subsatzes von Worten innerhalb eines Fensters von vorselektierter Breite vorausberechnet und gespeichert werden. Wenn es mit einer Anfrage konfrontiert ist, berechnet das System den Satz von Schlüsseln für die Anfrage und verwendet sie dazu, den Satz von Schlüsseln zu suchen und zu finden, der für alle Redewendungen in allen Büchern berechnet wurde. Mit anderen Worten werden, anstatt die Redewendungen direkt miteinander zu vergleichen, statt dessen deren "Repräsentanten" verglichen, wiederum in einer linearen Weise.

Eine Situation, die analog zu der obigen Suchaufgabe nach ähnlichen Redewendungen ist, existiert im Fall der Suche nach ähnlichen Strukturen in Datenbanken dreidimensionaler molekularer Information. Die folgende Darstellung repräsentativer Techniken ist dazu gedacht, bei der Identifizierung der Gemeinsamkeiten und der Unterschiede von zuvor vorgeschlagenen Vorgehensweisen zu helfen.

Die verschiedenen Techniken, die über die Jahre hinweg vorgeschlagen wurden, um dreidimensionale molekulare Datenbanken zu durchsuchen, unterscheiden sich im Grundsatz hinsichtlich ihrer Definition und ihrer Verwendung des Ähnlichkeitsmaßes $d(\dots)$, das oben eingeführt wurde. Sind eine Testverbindung C und eine Datenbank D gegeben, berechnet das Ähnlichkeitsmaß $d(\dots)$ das Ausmaß, in dem C und ein gegebenes Mitglied von D ähnlich sind. Die Werte, die dadurch erzeugt werden, daß C mit jedem der Mitglieds-moleküle von D verglichen wird, erzeugen eine "Markierung", die nachfolgend dazu verwendet werden kann, Antworten von Kandidaten in der Reihenfolge abnehmender Qualität zu ordnen.

Zum Beispiel wird in dem Verfahren der "Abbildung von Atomen" unter Verwendung des Ergebnisses von paarweisen Vergleichen der Zeilen von den Abstandsmatrizen zweier Moleküle der Tanimoto-Koeffizient berechnet. Dieser Koeffizient wird als Eingabe in eine intermolekulare Ähnlichkeitsmatrix verwendet. Diese Matrix wird in Verbindung mit einem aufwendigen Algorithmus verwendet, um den Grad der Ähnlichkeit zwischen den zwei Molekülen festzustellen. Die Berechnung wird für alle Kombinationen zwischen einem Anfragemolekül C und jedem der Moleküle in der Datenbank D wiederholt. Wie es mit aufwendigen Algorithmen der Fall ist, gibt es keine Garantien, daß der Algorithmus alle richtigen Lösungen findet. Die Vorgehensweise stellt berechnungsmäßig hohe Anforderungen und skaliert nicht gut mit der Datenbankgröße.

Bei dem Verfahren der "Cliques-Detektion" wird eine Anzahl verschiedener Orientierungen für jedes der Moleküle in der Datenbank erzeugt, bevor es mit dem Anfragemolekül C verglichen wird. Jede der Orientierungen wird dann über C gelegt und markiert, basierend auf dem Vorhandensein oder Nichtvorhandensein von Datenbankatomen in der Nachbarschaft eines Atoms von C. Alle Orientierungen, die dazu führen, daß weniger Treffer als maximal möglich stattfinden, werden ausgeschieden. Die Suche fährt dann mit dem nächsten Molekül der Datenbank fort. An jeglichem Punkt während der Suche werden die letzten n besten Markierungen behalten. Diese Technik steht im Kern des molekularen Mosaik-Modelliersystems.

Bei anderen bekannten Techniken werden die molekularen Strukturen als Verknüpfungstafeln dargestellt und somit als Graphen angesehen. Die Vertices jedes derartigen Graphen entsprechen den Atomplätzen des Moleküls. Wenn zwischen zwei gegebenen Atomplätzen eine Bindung existiert, dann weist der entsprechende Graph eine Ecke auf, welche die relevanten Knoten verbindet. Wenn jedes Molekül der Datenbank D durch einen Graph dargestellt wird, kann man eine Suche nach ähnlichen Substrukturen durch Verwenden von Subgraphisomorphismus-Algorithmen ausführen. Wie bereits oben erwähnt wurde, ist das Problem des Subgraphisomorphismus NP-vollständig, und somit existiert kein effizienter Algorithmus. Eine neuere Arbeit verglich eine Anzahl verschiedener Subgraphisomorphismus-Algorithmen und legt Beweise für die Nützlichkeit eines rückverfolgenden Suchalgorithmus dar, der mit der "Verfeinerungsprozedur"-Heuristik verbessert ist. Graphentheoretische Ergebnisse werden außerdem dazu verwendet, Ähnlichkeitsfunktionen zum Vergleichen von molekularen Fragmenten (Substrukturen) zu entwickeln.

Eine Variation des obigen Schemas beginnt durch Blockbildung aller Moleküle in der Datenbank D in mehrere Blöcke. In diesem Fall wird das Ähnlichkeitsmaß $d(\dots)$ zuerst dazu verwendet, intermolekulare Ähnlichkeiten für alle die Paare zu berechnen, die durch Moleküle in D gebildet werden können. Nachfolgend wird ein Blockbildungsschritt dazu verwendet, die verschiedenen Moleküle in Blöcke zu gruppieren, basierend auf den Werten, die durch die paarweisen Vergleiche erzeugt wurden. Bei Vorlegen eines Anfragemoleküls C klassifiziert diese Vorgehensweise C durch Identifizieren des Blockes, zu dem C gehört. Die Moleküle der Datenbank

D, die am besten mit dem Anfragemolekül C zusammenpassen, werden diesem Block ebenso wie dem Nachbarblock (oder den Nachbarblöcken) entnommen.

Bisher war die Annahme, daß die betrachteten Moleküle starre dreidimensionale Strukturen sind. Dies ist jedoch meistens nicht der Fall. Üblicherweise besitzen Moleküle mehrere interne drehbare Bindungen und sind somit in der Lage, ein Kontinuum von Konformationen, d. h. dreidimensionale Konfigurationen, anzunehmen. Hin und wieder können sterische Randbedingungen oder Energiebetrachtungen die Anzahl von Wahlmöglichkeiten begrenzen.

Werden die Moleküle einer Datenbank als starr behandelt, erleichtert dies die Suche in 3D-Datenbanken auf Kosten des Ausscheidens großer Mengen wahrer Kandidaten: Wenngleich die gespeicherte Konformation eines Moleküls möglicherweise nicht die (das) betrachtete pharmakophore Struktur/Modell zeigt, kann eine andere Konformation des gleichen Moleküls biologisch aktiv sein. Somit öffnet die Konformationsflexibilität von molekularen Strukturen einen breiten Bereich von Möglichkeiten bei der Suche nach potentiellen Liganden. Gleichzeitig wird dadurch jedoch der Suchkomponente der herkömmlichen Vorgehensweisen eine ernsthaftige Bürde auferlegt.

Ist eine Datenbank D von Molekülen gegeben, erfordert eine direkte Vorgehensweise, die es jedem beliebigen Suchalgorithmus erlaubt, konformationsflexible Suchen in D durchzuführen, das Speichern aller Konformationen von jedem der Moleküle in D. In der Praxis wird in Betracht eines Kontinuums möglicher Konformationen statt dessen eine große Anzahl repräsentativer Konformationen gespeichert. Die Auswirkungen einer derartigen Vorgehensweise sind offensichtlich: Die resultierenden Datenbanken haben überwältigende Größen, und es sind sehr lange Suchzeiten notwendig. Eine alternative Vorgehensweise zum Speichern aller möglichen Konformationen beinhaltet das Speichern jedes Moleküls in nur einer (oder einer Handvoll von) Konformation(en). Zum Beispiel verwendet das Concord-3D-System einen Satz von Regeln, um eine einzelne Konformation zu erzeugen, wobei die Verknüpfungstafeln des Moleküls verwendet werden. Diese Vorgehensweisen gehören im wesentlichen zu einer Klasse von Verfahren, welche die Flexibilität in die Datenbank legt.

In einer analogen Weise wendet eine Variante dieses Verfahrens einen Satz von Regeln an (bestimmt mittels Durchführen einer systematischen Konformationsanalyse von Ketten verschiedener Kombinationen von sechs Hauptgerüstatomen), um den Konformationsraum zu untersuchen und lediglich bestimmte Torsionswinkel für jede drehbare Bindung zu behalten: Es wird ein Satz von "Niederenergie"-Konformationen zusammen mit ihren entsprechenden "Masken" erzeugt. Diese Masken werden nachfolgend während der aktuellen Durchsuchung der Datenbank verwendet. Bei einer verwandten Vorgehensweise wird eine große Anzahl von Konformationen des Moleküls der Datenbank während der Durchsuchung erzeugt und mit der pharmakophoren Struktur verglichen. Dies ist eine berechnungsmäßig aufwendige Vorgehensweise, und jegliche Versuche (durch die Verwendung von Heuristiken), diese Belastung zu reduzieren, haben eine direkte Einwirkung auf die Qualität der erzeugten Ergebnisse: sonst zutreffende Gegenstücke werden nun übersehen. Beide dieser Verfahren sind für eine Klasse von Techniken repräsentativ, welche die Konformationsflexibilität in die Suche verlegen.

Es gibt außerdem eine dritte Vorgehensweise, bei der die Flexibilität in die Anfrage selbst verlegt wird. Die Anfrage kombiniert in diesem Fall sowohl starre als auch flexible Komponenten und wird mittels Durchsuchen einer Datenbank von Verbindungen mit "bekannter" Aktivität iterativ verfeinert, bis die gewünschte Selektivität erzielt wird. Steht einmal die endgültige Anfrage zur Verfügung, wird sie dazu verwendet, eine Datenbank von Verbindungen mit "unbekannten" Aktivitäten zu durchsuchen, um potentielle Verbindungen zu identifizieren.

Die erfolgreicher Suchtechniken greifen das Problem der Konformationsflexibilität in einer berechnungsmäßig herausfordernden Weise an) die Schlußfolgerung früherer Arbeiten war, daß ein flexibles dreidimensionales Suchen unter Verwendung der von Clark et al. entwickelten Vorgehensweise als ein Minimum eine 100fache Verlangsamung gegenüber dem starren Vergleichsfall zur Folge hat. Diese Verlangsamung scheint typisch für und unabhängig von der aktuellen Technik zu sein, die verwendet wird.

In einer Vergleichsstudie von Haraki et al. wurde gezeigt, daß ein Vergrößern einer Datenbank mit Mehrfachkonformationen eines gegebenen Moleküls im allgemeinen die Leistungsfähigkeit eines Suchalgorithmus steigert. Die gleiche Studie stellte jedoch auch fest, daß die resultierende Effektivität stark von dem Verfahren abhängig ist, das zur Erzeugung der verschiedenen Konformationen, die zu der Datenbank hinzuzufügen sind, verwendet wird.

Als eine Alternative zur mehrfachen Aufnahme eines Moleküls in der Datenbank wird im "Diskrepanzraum" ein bestimmter Typ von Minimierung ausgeführt. Diese Vorgehensweise ist viel schneller, sie erfordert jedoch bestimmte Beziehungen zwischen der Anzahl von strukturellen Randbedingungen und der Anzahl von existierenden drehbaren Bindungen; des weiteren übernimmt sie alle Probleme von nicht linearen Optimierungs-Vorgehensweisen.

Noch ein weiterer Typ von Technik versucht im wesentlichen, ein starres Andocken lediglich auf den starren Subteilen des Moleküls durchzuführen und dann die Kompatibilität der verschiedenen angedockten Teile in einer Nachbearbeitungsphase zu überprüfen. Diese Technik stellt im allgemeinen berechnungsmäßig hohe Anforderungen.

Eine Anzahl von Abkürzungen in Form von Such-Heuristiken wurde eingeführt, um einiges an berechnungsmäßiger Belastung zu vermindern, jedoch nicht ohne nachteilige Auswirkung auf die Qualität der erzeugten Ergebnisse. Um ein Gegengewicht zu dieser Aussage zu bilden, sind diese Heuristiken von einer allgemeineren Anwendbarkeit und können auch in dem Fall verwendet werden, in dem Konformationsflexibilität keiner der Parameter des Problems ist.

Insbesondere gibt ein gewisser Stand der Technik eine sehr gründliche Darstellung an und führt eine Vergleichsstudie einer Anzahl von Deskriptoren zum Zwecke einer Datenbanksichtung aus. Die Deskriptoren decken einen großen Bereich an Eigenschaften der Moleküle in der Datenbank ab: physikalische, chemische, geometrische ebenso wie verschiedene Kombinationen derselben. Dies Unterscheidungsfähigkeit einiger der

vorgeschlagenen Deskriptoren ist ermutigend, die Ergebnisse wurden jedoch unter Verwendung einer kleinen Datenbank mit nur einigen wenigen tausend Verbindungen erzielt.

Dazu in Beziehung stehende Arbeiten führten ein zweistufiges Verfahren ein, das im wesentlichen die Gestalt charakterisiert, ohne die Notwendigkeit zur Prüfung einer Vielzahl von Andockorientierungen. In der ersten Stufe wird für jedes der Moleküle in der Datenbank durch Setzen geeigneter Bits in einem Bit-Vektor eine Zahl mit 2048 Bit erzeugt. Die zu setzenden Bits werden basierend auf einer 32-Bit-Codierung jedes Dreiecks ausgewählt, das durch drei Atomplätze in dem Molekül gebildet wird. Klarerweise enthält die Zahl geometrische Charakteristika, die für jedes Molekül spezifisch sind; aufgrund der Art und Weise, in der sie erzeugt wird, ist die Darstellung jedoch nicht eindeutig. Während der zweiten Stufe wird eine ähnliche 2048-Bit-Zahl für das Testmolekül erzeugt und mit jedem der gespeicherten Signaturen verglichen. Für jene Moleküle, deren Signaturen einen Schwellwert überschreiten, werden wiederum Triplets von Atomplätzen gebildet und mit den Triplets in dem Testmolekül zwecks Überschneidung verglichen. Wenngleich das Verfahren die relative Orientierung zwischen einem Kandidaten und dem Testmolekül nicht gewinnt, erweist es sich als Sichtungsschritt adäquat.

Im gesamten beschriebenen Stand der Technik skalieren die Techniken entweder nicht gut mit der Größe der Datenbank (aufgrund der Notwendigkeit für eine serielle Abtastung und Bearbeitung aller Einträge) oder verwerten die durch die drehbaren Bindungen auferlegten Randbedingungen, um das Ausmaß der Suche zu begrenzen, nicht vollständig.

Anders als bei den Techniken, die eine lineare Abtastung der Datenbank D erfordern, basieren Hash-Techniken auf der Identifikation bestimmter invarianter Deskriptoren (Indizes), die dazu verwendet werden können, eine partielle Darstellung, sagen wir eines Moleküls, in einer Suchtabelle zu speichern. Kompatible Moleküle können durch Berechnen der Indizes von einer Testeingabe wiedergewonnen werden, indem die partielle Darstellung aus der Suchtabelle zurückgeholt und das Ergebnis direkt integriert wird, wodurch die Notwendigkeit eliminiert wird, die gesamte Datenbank für einen oder mehrere Vergleiche abzutasten. Für Moleküle können Indizes durch Verwenden von Atomtupeln (z. B. Triplets) mit Atomeigenschaften oder Tupeln mit kleinen Oberflächenstücken gebildet werden, die mit ihren Normalen und den chemischen Eigenschaften an der Oberfläche verknüpft sind (zwei unabhängige Stücke sind in diesem Fall ausreichend).

In früheren Arbeiten wurde argumentiert, daß die Verwendung von Indizes hochdimensionaler Natur (mit einer großen Anzahl von einzelnen Werten) entscheidend für das richtige Verhalten dieser Techniken ist, wenn die Ausdehnung der Datenbank groß wird. Zwei Hauptpunkte tragen zu diesem sehr allgemeinen Ergebnis bei. Erstens sind Fächer in einer Suchtabelle mit einem größeren Satz von Fächern im Durchschnitt weniger besetzt. Und zweitens kann eine gröbere Quantisierung entlang jeder der Indexdimensionen verwendet werden, womit die Wahrscheinlichkeit der Gewinnung des gleichen Index während des Wiederauffindens ähnlicher Entitäten erhöht wird.

Anders als bei auf Abtastung basierenden Techniken hat jedoch die Klasse von Hash-Technik-Algorithmen die Speicheranforderungen erhöht. Insbesondere die verschiedenen Stufen des Algorithmus leiten ihre Geschwindigkeit aus dem Vorausberechnen von Ergebnissen und dem Speichern derselben in geeignet aufgebauten Suchtabellen ab. Diese Vorausberechnung kann off-line durchgeführt werden, erfolgt lediglich einmal, und die Ergebnisse werden auf Platte gespeichert und verwendet, wenn notwendig. Die Hash-Technik-Vorgehensweise schafft im wesentlichen Raum für die Berechnung; angesichts einer Verminderung der Kosten langsamer Speicher wird der Kompromiß zunehmend vertretbar und vernünftig.

AUFGABEN DER ERFINDUNG

Eine Aufgabe dieser Erfindung ist ein verbessertes Rechnersystem und Verfahren zur Bestimmung jener Moleküle aus einer Datenbank D, die ein oder mehrere Moleküle, die Substrukturen beinhalten, die Substrukturen von einem oder mehreren Testmolekülen C entsprechen, zusammen mit dem Satz von starren Transformationen (d. h. starre Rotationen und Translationen) enthält, der bewirkt, daß jedes dieser Moleküle am besten mit dem (den) Testmolekül(en) C überlappt ("beste Übereinstimmung"), das beschriebene System und Verfahren können dies erreichen, auch wenn die Moleküle in der Datenbank Gruppen von Atomen enthalten, die frei um jegliche kovalente Bindungen herum rotieren können, die möglicherweise in dem Molekül vorhanden sind (Torsionsflexibilität).

ZUSAMMENFASSUNG DER ERFINDUNG

Dieses System und dieses Verfahren identifizieren Moleküle und/oder molekulare Substrukturen in einer Datenbank D, die ähnlich oder identisch mit einem oder mehreren Testmolekülen und/oder Substrukturen und/oder Teilen von Substrukturen dieser Testmoleküle sind. Das System und das Verfahren legen außerdem den Satz starrer Transformationen (d. h. starrer Rotationen und Translationen) fest, der bewirkt, daß jedes der Moleküle, die in D identifiziert sind, am besten mit dem (den) Testmolekül(en) überlappt. Dies wurde oben als "beste Übereinstimmung" bezeichnet. Anders als bei Techniken, die im Stand der Technik enthalten sind, ist die Festlegung der geeigneten starren Transformationen weder das Ergebnis irgendeiner erschöpfenden Suche im Raum möglicher Konformationen, noch irgendeiner anderen äquivalenten Suchprozedur. Statt dessen tritt sie gleichzeitig mit der Festlegung der Identitäten jener Moleküle in D auf, die Substrukturen enthalten, die ähnlich oder identisch mit Substrukturen sind, die in dem (den) Testmolekül(en) enthalten sind. Wiederum ist aufgrund der auf Hash-Techniken basierenden Natur der Erfindung keine lineare Abtastung der Moleküle in D notwendig.

Die Erfindung verwendet einen Referenzspeicherprozeß, um eine Datenstruktur zu bevölkern, so daß die Datenstruktur alle molekularen Strukturen und/oder Substrukturen in der Datenbank enthält, die gemäß

Eigenschaften von Tupeln klassifiziert sind. In einer bevorzugten Ausführungsform werden die Tupel von Plätzen (z. B. Atomplätzen) der molekularen Strukturen (Substrukturen), abgeleitet, die gewählt sind, um die Tupel zu erzeugen, und die Eigenschaften sind geometrische (und andere) Informationen, die sich auf die gewählten Tupel beziehen. Die Eigenschaften werden dazu verwendet, Indizes in einer Datenstruktur zu definieren, die mit invarianten Vektorinformationen (Vektorinformation genannt) verknüpft sind, die zu den Molekülen der Datenbank D gehören. Zum Beispiel können die invarianten Vektoren drehbare Bindungen in Referenzmolekülen in der Datenbank D repräsentieren. Diese invarianten Vektoren (z. B. drehbare Bindungen) werden in schiefwinkligen lokalen Koordinatensystemen dargestellt, die von Tupeln erzeugt werden, die von starren molekularen Substrukturen abgeleitet sind, an welche der (die) Vektor(en) angebunden ist (sind). Diese Darstellungen sind invariant bezüglich der Rotation und Translation von molekularen Strukturen und/oder der Rotation von Substrukturen um die angebundene(n) drehbare(n) Bindung(en). Demgemäß können invariante Vektorinformationen, die zu Molekülen in der Datenbank gehören, bezüglich der Tupeleigenschaften dadurch klassifiziert werden, daß die invarianten Vektorinformationen an Stellen (Vektorfeldern) der Datenstruktur gespeichert werden, die mit dem von dem jeweiligen Tupel abgeleiteten Index verknüpft sind. Sobald die Datenstruktur bevölkert ist, erzeugt ein Vergleichsprozeß einen oder mehrere Tupel, schiefwinklige lokale Referenzkoordinatensysteme und Indizes (Testkoordinatensystemtupelindizes genannt) für die Struktur (Substrukturen) eines Testmoleküls unter Verwendung der gleichen Technik, die zur Bevölkering der Datenstruktur verwendet wurde.

Der Testkoordinatensystemtupelindex wird dazu verwendet, auf die invariante Vektorinformation zuzugreifen, die sich in dem Vektorfeld des Datenstrukturindex befindet, der mit dem Testkoordinatensystemindex zusammenpaßt. Es wird ein Zählwert der Häufigkeit des Zusammenpassens der Vektorinformationen (Indizes) von molekularen Strukturen (Substrukturen und/oder Teilen) in der Datenbank mit den Testkoordinatensystemtupelindizes, die für das Testmolekül erzeugt wurden, um zu bestimmen, welche molekularen Strukturen (Substrukturen und/oder Teile) identisch oder ähnlich zu jenen in der Datenbank passen, und der notwendigen starren Transformationen behalten, die zugehörige Substrukturen in Übereinstimmung miteinander bringen. In einer bevorzugten Ausführungsform wird die Kombination von gezählten Vektorinformationen und den notwendigen starren Transformationen von Kandidaten ausgeschieden, wenn sie nicht durch eine oder mehrere andere Substrukturen bestätigt ist.

KURZBESCHREIBUNG DER ZEICHNUNGEN

Die vorstehenden und weitere Aufgaben, Aspekte und Vorteile der Erfindung werden aus der folgenden detaillierten Beschreibung von bevorzugten Ausführungsformen der Erfindung unter Bezugnahme auf die Zeichnungen besser verständlich werden, die wie folgt beschrieben sind:

Fig. 1 ist ein Blockdiagramm eines Rechnersystems, das die vorliegende Erfindung verkörpert.

Fig. 2A ist eine Darstellung einer molekularen Struktur, die starre Substrukturen von Atomgruppen in dem Molekül, die Rotationsnatur von zwei typischen Drehbindungen zwischen starren Substrukturen, ein globales Koordinatensystem, ein schiefwinkliges lokales Koordinatensystem, ein "Koordinatensystem-Tupel", welches das schiefwinklige lokale Koordinatensystem definiert, und eine Darstellung von zwei invarianten Vektoren, die zwei Punktpaare auf zwei oder mehr starren Substrukturen verbinden, sowie eine erste Konformation der molekularen Struktur zeigt.

Fig. 2B ist eine Darstellung der molekularen Struktur von Fig. 2A, die ein "magisches Koordinatensystem" zeigt, das mit einer der starren Substrukturen verknüpft ist.

Fig. 2C ist eine Darstellung, die eine zweite molekulare Konformation der molekularen Struktur, das globale Koordinatensystem, das schiefwinklige lokale Koordinatensystem und die invarianten Vektoren von Fig. 2A zeigt.

Fig. 3 ist eine Sequenz von Zeichnungen, die zeigen, wie Platzsätze K-O definiert sind (Fig. 3A), wie Dummy-Stellen Du definiert sind und dann verwendet werden (Fig. 3B) und wie Tupel (Fig. 3B bis 3E) durch Wählen eines Satzes von einem oder mehreren Atomplätzen und/oder Dummy-Plätzen Du von der molekularen Struktur definiert sind.

Fig. 4 ist ein Blockdiagramm einer Datenstruktur, die einen Index, der zu einem Tupel gehört, mit einer Vektorinformation verknüpft, die der Darstellung von zwei oder mehr invarianten Vektoren entspricht, die an jedes der schiefwinkligen lokalen Koordinatensysteme des Tupels, das den Index erzeugt, angebunden sind.

Fig. 5, welche die Fig. 5A und 5B umfaßt, ist ein Flußdiagramm, das die Schritte der Bevölkering der Datenstruktur von Fig. 4 zeigt, damit sie strukturelle Informationen und andere Informationen über ein oder mehrere Referenzmoleküle enthält.

Fig. 6, welche die Fig. 6A, 6B und 6C umfaßt, ist ein Flußdiagramm eines bevorzugten Verfahrens, das die Schritte der Bestimmung, (a) welche Referenzmoleküle in der Bibliothek (Datenbank D), die ein oder mehrere Moleküle enthält, einem Testmolekül für einen ausgewählten Satz einer oder mehrerer molekularer Eigenschaften ähnlich sind (= mit diesem zusammenpaßt) und (b) des Satzes von starren Transformationen zeigt, die das Testmolekül mit den in (a) identifizierten Referenzmolekülen in Übereinstimmung bringt.

Fig. 7 ist ein Blockdiagramm einer Abstimmtable, die dazu verwendet wird, die relative Frequenz (Häufigkeitswerte) von:

(a) den Identitäten jener Moleküle und/oder molekularen Substrukturen in der Datenbank D, die mit dem Testmolekül für einen gegebenen Satz von molekularen Eigenschaften zusammenpassen, und
(b) den starren Transformationen zu bestimmen, die das Testmolekül mit den in (a) identifizierten Referenzmolekülen in Übereinstimmung bringen.

DETAILLIERTE BESCHREIBUNG DER ERFINDUNG

Nunmehr beziehend auf die Zeichnungen und spezieller auf Fig. 1, ist dort die Blockdiagrammdarstellung einer allgemeinen Rechner-Hardwareumgebung 100 gezeigt. Dieser Rechner 100 kann einer aus der International Business Machines Corporation (IBM) Personal System/2 (PS/2) Familie von Personalcomputern, ein RISC System/6000 oder ein Power Parallel System (SP/x) sein. Das System 100 beinhaltet eine oder mehrere Zentralprozessoreinheiten (CPU) 10, die der x86-Architektur von Intel entsprechen oder aus einem Mikroprozessor mit reduziertem Befehlssatz bestehen können. Die CPU 10 ist an einen Systembus 12 angebunden, mit dem ein Schreib-/Lese-Speicher und/oder ein Speicher mit wahlfreiem Zugriff (RAM) 14 verbunden ist, die einen oder mehrere Cache-Speicher, einen Festspeicher (ROM) 16, einen Eingabe-/Ausgabe-Adapter 18 und einen Benutzerschnittstellen-Adapter 22 umfassen können. Das RAM 14 sorgt für eine temporäre Speicherung eines oder mehrerer Anwenderprogramme 40, die einen Code und/oder Daten enthalten, während das ROM 16 typischerweise den grundlegenden Eingabe-/Ausgabesystem(BIOS)-Code beinhaltet. Der E/A-Adapter 18 ist mit einem oder mehreren Speichereinheiten mit direktem Zugriff (DASDs) verbunden, die hier als ein Floppy-Laufwerk 19, ein Festplatten-Laufwerk 20 und ein CD-ROM 21 dargestellt sind.

Das Festplatten-Laufwerk 20 speichert typischerweise das Rechnerbetriebssystem (OS), wie das OS/2-Betriebssystem von IBM, und verschiedene Anwenderprogramme, Daten und/oder Datenbanken 50, von denen jede über den Systembus 12 selektiv in das RAM 14 geladen werden kann. Mit dem Benutzerschnittstellen-Adapter 22 sind eine Tastatur 24, eine Maus 26 und/oder andere Benutzerschnittstelleneinheiten (nicht gezeigt) verbunden.

Das System 100 kann außerdem eine Anzeige 38 umfassen, die hier als Kathodenstrahlröhren(CRT)-Anzeige dargestellt ist, die jedoch auch eine Flüssigkristallanzeige (LCD) oder eine andere geeignete Anzeige und/oder eine graphische Benutzerschnittstelle (GUI) sein kann. Die Anzeige 38 ist über einen Anzeigeadapter 36 mit dem Systembus 12 verbunden. Ein Multimedia-Adapter 34, wie ein ActionMedia 11 Display Adapter von Intel Corporation, kann ebenfalls mit dem Bus 12 sowie mit einem Mikrophon 32 und einem Lautsprecher 28 verbunden sein. Der Multimedia-Adapter 34 wird von einer geeigneten Software, wie Multimedia Presentation Manager/2, unterstützt. Diese Systeme 100 und Äquivalente dieser Systeme sind dem Fachmann allgemein bekannt.

Personal System/2, PS/2, OS/2, RISC System/6000, Power Parallel System, SP/x und IBM sind Handelsmarken der International Business Machines Corporation.

Einige der Anwenderprogramme 40 werden unten als Prozeßverfahren beschrieben. Molekulare Datenbanken 50, die ebenfalls unten beschrieben sind, werden typischerweise in den Speichereinheiten gespeichert, z. B. dem Festplatten-Laufwerk 20.

Fig. 2A ist eine Darstellung einer molekularen Struktur 200, die starre Substrukturen (210, 220, 230) von Atomgruppen in dem Molekül 200, die Drehnatur 215 von zwei typischen drehbaren Bindungen (218a, 218b) zwischen starren Substrukturen (210, 220) beziehungsweise (220, 230), ein globales Koordinatensystem 235, ein schiefwinkliges lokales Koordinatensystem 245 und ein "Koordinatensystem-Tupel" zeigt, welches das schiefwinklige lokale Koordinatensystem 245 definiert. Außerdem ist in Fig. 2A die Darstellung von zwei invarianten Vektoren gezeigt: (a) dem Vektor 238, der zwei Punkte (D, G) auf einer oder mehreren starren Substrukturen (210, 220) verbindet, und (b) dem Vektor 248, der zwei Punkte (O, Q) auf einer oder mehreren starren Substrukturen (220, 230) verbindet. Schließlich zeigt Fig. 2A eine erste Konformation 200 der betrachteten molekularen Struktur.

Unten sind einige, der Ausdrücke, die im Text eingehend verwendet werden, definiert und klargestellt.

Eine molekulare Struktur (200, 250) ist ein Satz von Atomen (z. B. A-S), die über chemische Bindungen, typischerweise MO, miteinander verbunden sind. (Bindungen sind durch Buchstabenpaare bezeichnet, die den zwei durch die Bindung verbundenen Atomen entsprechen.) Im allgemeinen ist die molekulare Struktur 200 typischerweise durch einen Satz von Koordinaten für die von den verschiedenen Atomen besetzten Plätze definiert. Zum Beispiel definieren die Koordinaten (x, y, z) die Position von Atom O in dem globalen (Labor-) Koordinatensystem 235. Das globale Koordinatensystem 235 wird für die Zwecke der folgenden Analyse als fest und konstant angenommen.

Des weiteren definiert eine Liste der chemischen Bindungen, welche die Plätze miteinander verbinden, z. B. MO, ebenfalls die molekulare Struktur 200. Den verschiedenen Plätzen (A-S) der molekularen Struktur 200 und/oder den jeweiligen Atomen, die diese Plätze in dem globalen Koordinatensystem 235 besetzen, werden typischerweise Bezeichnungen (zum Beispiel: eine Zahl), zugeordnet, die sie voneinander unterscheiden. Für unsere Zwecke werden wir austauschbar einen Buchstaben (z. B. A-S) und/oder eine Zahl verwenden, um das Atom und/oder den Platz, an dem sich das Atom in dem globalen Koordinatensystem 235 befindet, zu bezeichnen.

Schließlich wird zusätzlich zu der Liste der Koordinaten der Plätze und der Liste von chemischen Bindungen eine Liste der Atomtypen (z. B. N, C, O, H etc.) für jedes der Atome bereitgestellt, welche die verschiedenen Plätze der molekularen Struktur besetzen.

Es ist zu erwähnen, daß gelegentlich eine molekulare Struktur unter Verwendung der Liste von chemischen Bindungen und der Liste von Atomtypen für jedes der Atome, die an der Struktur beteiligt sind, spezifiziert wird. Eine molekulare Struktur, die in einer derartigen Weise definiert wurde, ist auf keinen Fall unbestimmt: tatsächlich können die Koordinaten der Atomplätze aus der gegebenen Information unter Verwendung einer Vielzahl üblicher Verfahren gewonnen werden.

Eine Bindung, MO, repräsentiert eine chemische Verbindung zwischen zwei Atomen (M, O) in der molekularen Struktur 200. Bindungen sind typischerweise in Form der Bezeichnungen definiert, die mit den zwei Atomplätzen verknüpft sind, welche die Bindung verbindet.

Einige der Bindungen in einem gegebenen Molekül können drehbar sein und somit für eine Torsionsflexibilität 215 sorgen: die starren Substrukturen (210, 220) sind an den zwei Endpunkten einer drehbaren Bindung 218a angebunden und können daher relativ zueinander rotieren 215. In einer ähnlichen Weise sind die starren Substrukturen (220, 230) an den zwei Endpunkten einer drehbaren Bindung 218b angebunden und können daher relativ zueinander rotieren 215.

Innerhalb jeder der drei starren Substrukturen (210, 220, 230) auf jeder Seite von drehbaren Bindungen (218a, 218b) sind Atome miteinander durch Bindungen verbunden, die eine derartige Torsionsflexibilität nicht erlauben (nicht drehbare Bindungen — AC, MO). Dies trifft im allgemeinen Fall nicht zu: es existieren molekulare Substrukturen, die eine scharnierartige Flexibilität zeigen, eine Behandlung dieser Substrukturen übersteigt jedoch den Umfang dieser Analyse.

Daher sind die starren Substrukturen (210, 220, 230) Strukturen aus einem oder mehreren Atomen, die über nicht drehbare Bindungen miteinander verbunden sind. Gruppen, die aus einem Atom, P, bestehen, das über Bindungen wie NP mit einem Satz von Atomen wie G, H, I, J, K, L, M, N und O verbunden sind, werden nicht als separate starre Substrukturen betrachtet, ungeachtet der Tatsache, daß die Bindung NP möglicherweise drehbar ist. Dies liegt daran, daß jegliche Rotation des Atoms P um die Bindung NP herum den Ort von P in dem globalen Koordinatensystem 235 nicht ändert. Des weiteren ändert jegliche Rotation des Atoms P um die Bindung NP herum den Ort von P bezüglich des Satzes von Atomen G, H, I, J, K, L, M, N und O nicht. Es ist außerdem zu erwähnen, daß starre Moleküle, d. h. jene Moleküle, die keine drehbaren Bindungen enthalten, als Moleküle mit einer starren Substruktur definiert werden können; in einem derartigen Fall ist das gesamte Molekül die Substruktur (210, 220, 230).

Es ist außerdem zu erwähnen, daß die Definition der Koordinaten (x, y, z) von drei oder mehr Atomen (Plätzen) (z. B. G-P einer gegebenen starren Substruktur 220) in dem globalen Koordinatensystem 235 genügt, um eine globale Position $O''O'$ und eine globale Orientierung ($O'x', O'y', O'z'$) für die starre Substruktur, z. B. 220, in dem globalen Koordinatensystem 235 zu definieren. Man beachte außerdem, daß der Satz von drei oder mehr Atomen (Plätzen), der die globale Position und Orientierung für die starre Substruktur, z. B. 220, definiert, das Atom (den Platz) D beinhalten kann, da die Rotation 215a um die drehbare Bindung 218a der starren Substruktur 210 herum in bezug auf die starre Substruktur 220 die Position des Atoms (dem Platz) D bezüglich der starren Substruktur 220 nicht ändert. Bei Verwenden eines ähnlichen Arguments kann der Satz von drei oder mehr Atomen (Plätzen), welche die globale Position und Orientierung für die starre Substruktur 220 definieren, das Atom (den Platz) Q beinhalten, da die Rotation 215b um die drehbare Bindung 218b der starren Substruktur 230 herum in bezug auf die starre Substruktur 220 die Position des Atoms (den Platz) Q bezüglich der starren Substruktur 220 nicht ändert. Es ist außerdem zu erwähnen, daß bei der Definition einer globalen Position und Orientierung für die starre Substruktur 210 der Satz von drei oder mehr Atomen (Plätzen) das Atom (den Platz) G zusätzlich zu den Atomen (Plätzen) A-F beinhalten kann. In einer analogen Weise kann bei der Definition einer globalen Position und Orientierung für die starre Substruktur 230 der Satz von drei oder mehr Atomen (Plätzen) das Atom (den Platz) O zusätzlich zu den Atomen (Plätzen) Q-S beinhalten.

Demzufolge besitzt der Vektor 238 (bzw. 248), der unten definiert ist, eine feste Position und Orientierung bezüglich entweder der Substruktur 210 oder 220 (bzw. 220 oder 230), welche die drehbare Bindung 218a (bzw. 218b) verbindet. Dies geschieht, weil sich die Position und Orientierung der drehbaren Bindung 218a (bzw. 218b) bezüglich jeder starren Substruktur ungeachtet der Rotation in dem globalen Koordinatensystem 235 jeder der Substrukturen um die drehbare Bindung 218a (bzw. 218b) herum nicht ändern.

Im folgenden kann der Ausdruck starre Substruktur (210, 220, 230) austauschbar mit dem Ausdruck starre Gruppe (210, 220, 230) verwendet werden.

Wie unten erörtert wird, brauchen die Vektoren 238 (bzw. 248), die, wie bereits erwähnt, eine feste Position und Orientierung bezüglich entweder der starren Substruktur 210 oder 220 (bzw. 220 oder 230) besitzen, nicht in Form der drehbaren Bindung(en) 218a (bzw. 218b), die von der starren Substruktur ausgehen, definiert zu werden. Tatsächlich kann der Vektor 238 (bzw. 248) für eine gegebene starre Substruktur jeder Vektor sein, der so definiert werden kann, daß er bezüglich der starren Substruktur starr angeordnet ist.

Für den Augenblick wird der Vektor 238 (bzw. 248) mit Hilfe der drehbaren Bindung 218a (bzw. 218b) definiert: zum Beispiel fallen die Größe und die Richtung des Vektors 238 (bzw. 248) mit jenen der Bindung 218a (bzw. 218b) zusammen. Es wird angenommen, daß die Konvention für die Richtung diejenige von der Substruktur mit niedrigerer (höherer) Numerierung (210, 220) (bzw. 220, 230) zu der Substruktur mit der höheren (niedrigeren) Numerierung (210, 220) (bzw. 220, 230) ist, konsistent für alle der einen oder mehreren analysierten molekularen Strukturen 200. Eine alternative Konvention für die Richtung basiert auf den Bezeichnungen der Atome (Plätze) an den Endpunkten einer drehbaren Bindung: Es wird angenommen, daß die Richtung von dem Atom (dem Platz) mit der niedrigeren (höheren) Numerierung zu dem Atom (dem Platz) mit der höheren (niedrigeren) Numerierung verläuft, konsistent für alle der einen oder mehreren analysierten molekularen Strukturen 200.

Eine gegebene molekulare Struktur 200 kann mehr als eine drehbare Bindung 218a (bzw. 218b) enthalten und kann somit über Drehungen um diese Bindungen 218a (bzw. 218b) herum jede beliebige einer möglicherweise unendlichen Anzahl von Konfigurationen (200, 250) annehmen. Molekulare Strukturen 200 mit einer oder mehreren drehbaren Bindungen 218a (bzw. 218b) werden als "konformationsflexible" molekulare Strukturen oder "konformationsflexible" Moleküle bezeichnet.

Es ist zu erwähnen, daß die molekulare Struktur 250 eine weitere Konformation der molekularen Struktur 200 ist (und umgekehrt), da sie die gleiche molekulare Struktur ist, wobei (a) ihre starren Substrukturen (210, 220) um die drehbare Bindung 218a relativ zueinander gedreht sind 215a und (b) ihre starren Substrukturen (220, 230) um die drehbare Bindung 218b relativ zueinander und unabhängig von der Rotation in (a) gedreht sind 215b.

Alternativ kann eine gegebene molekulare Struktur, 200 keine drehbaren Bindungen (218a, 218b) enthalten,

und sie wird dann als eine "starre" molekulare Struktur oder ein "starres" Molekül bezeichnet.

Zusätzlich zu der Konformationsflexibilität einer molekularen Struktur (200, 250) durch Rotationen um ihre drehbaren Bindungen (218a, 218b) herum kann die gesamte molekulare Struktur (200, 250) außerdem mit drei Freiheitsgraden rotieren 290 und sich mit drei Freiheitsgraden in dem globalen Koordinatensystem 235 verschieben 295.

Zusätzlich zu dem globalen Koordinatensystem 235 können außerdem "lokale" Koordinatensysteme 245 durch geeignetes Auswählen eines kleinen Satzes von Atomplätzen (z. B. I, K, H) in der molekularen Struktur (200, 250) erzeugt werden. Zum Beispiel können, wenn die drei Atomplätze I, K und H (die so gewählt sind, daß sie nicht kollinear sind) in der molekularen Struktur (200, 250) gegeben sind, die Vektoren $i = I \rightarrow H$ und $j = I \rightarrow K$ erzeugt werden. Da angenommen wird, daß die drei Punkte nicht kollinear sind, ist das Kreuzprodukt $k = i \times j$ der zwei Vektoren i und j wohldefiniert und senkrecht zu der Ebene, die durch die Vektoren i und j definiert ist. Die Einheitsvektoren u_1 , u_2 und u_3 entlang der Richtungen, die durch die drei Vektoren i , j beziehungsweise k definiert sind, definieren ein schiefwinkliges lokales Koordinatensystem 245. Dieses Koordinatensystem 245 wird als "schiefwinklig" bezeichnet, da die Einheitsvektoren i und j im allgemeinen Fall nicht orthogonal zueinander sind. Es ist jedoch möglich, daß die erzeugten schiefwinkligen lokalen Koordinatensysteme 245 aus Einheitsvektoren u_1 und u_2 bestehen, die orthogonal zueinander sind.

Es ist zu erwähnen, daß, wie oben beschrieben, ein schiefwinkliges lokales Koordinatensystem 245 durch Auswählen einer (oder beider) der Atomplätze, welche die drehbare Bindung 218a definieren, D oder G (oder D und G) und von zwei (oder einer) der restlichen Atomplätze der gegebenen Substruktur (210, 220) erzeugt werden kann. Zum Beispiel: ein schiefwinkliges lokales Koordinatensystem 245 für die Substruktur 210 kann durch Verwenden von einem der Atomplätze H, I, J, K, L, M, N, O, P und sowohl D als auch G definiert werden. Äquivalent kann ein schiefwinkliges lokales Koordinatensystem 245 für die Substruktur 220 durch Verwenden von zwei der Atomplätze H, I, J, K, L, M, N, O, P und genau einem von D, G definiert werden. In einer ähnlichen Weise können ein oder mehr schiefwinklige lokale Koordinatensysteme 245 für die Substruktur 210 gewählt werden. Eine analoge Beobachtung kann für die schiefwinkligen lokalen Koordinatensysteme 245 gemacht werden, die so erzeugt werden, daß sie Atomplätze von den molekularen Substrukturen (220, 230) und der drehbaren Bindung 218b beinhalten.

Das globale Koordinatensystem 235 ist von den schiefwinkligen lokalen Koordinatensystemen 245, die man erzeugen kann, verschieden, da die Position und Orientierung des schiefwinkligen lokalen Koordinatensystems 245 in dem globalen Koordinatensystem 235 variieren kann, wenn das jeweilige Molekül 200 starre Transformationen (Drehungen 290 und Translationen 295) erfährt. Die gleiche Aussage gilt, wenn eine starre Gruppe (210, 220, 230) in der molekularen Struktur (200, 250) relativ zu einer anderen starren Gruppe (210, 220, 230) um die drehbare Bindung 218a (bzw. 218b) herum rotiert 215a (bzw. 215b), welche die zwei starren Gruppen 210 und 220 (220 bzw. 230) verbindet.

Es ist zu erwähnen, daß die Vektoren 238 und 248 so definiert wurden, daß sie bezüglich der starren Substruktur 220 starr angeordnet sind, und somit auch das ausgewählte schiefwinklige lokale Koordinatensystem 245. Des weiteren befinden sich die Vektoren 238 und 248 stets in einer festen Position und Orientierung in bezug zueinander. Von diesem Punkt an werden die zwei Vektoren 238 und 248 als das "magische Vektorpaar" bezeichnet, das mit der molekularen starren Substruktur 220 verknüpft ist. Jede molekulare starre Substruktur 210, 220, 230 eines gegebenen Moleküls 200, 250 besitzt ein magisches Vektorpaar, das mit derselben verknüpft ist. Jedes der Elemente des magischen Vektorpaares (d. h. 238 und 248) wird als ein "magischer Vektor" bezeichnet. Wie bereits oben angedeutet, braucht ein magischer Vektor, der mit einer gegebenen Substruktur verknüpft ist, nicht notwendigerweise in Form einer der drehbaren Bindungen definiert zu sein, die von der betrachteten Substruktur ausgehen. Im folgenden kann der Ausdruck magisches Vektorpaar (bzw. magischer Vektor) austauschbar mit dem Ausdruck Referenzvektorpaar (bzw. Referenzvektor) verwendet werden.

Im allgemeinen Fall wird angenommen, daß die zwei Vektoren 238 und 248 nicht kollinear sind und somit ein Koordinatensystem erzeugen, das als das "magische Koordinatensystem" 255 bezeichnet wird. Es kann angenommen werden, daß das magische Koordinatensystem (in Fig. 2B gezeigt) seinen Ursprung O'' im Schwerpunkt der vier Endpunkte (zwei Anfangs- und zwei Endpunkte) der zwei magischen Vektoren hat; die Hauptachsen $O''x$ und $O''y$ des Koordinatensystems sind durch Translation der Vektoren 238 und 248 derart definiert, daß ihre Ursprünge mit dem Ursprung O'' zusammenfallen. Da angenommen wird, daß die zwei Vektoren 238 und 248 nicht kollinear sind, ist ihr Kreuzprodukt wohldefiniert und bestimmt die dritte Hauptachse $O''z$ des magischen Koordinatensystems. Da die Vektoren 238 und 248, die das magische Koordinatensystem definieren, nicht notwendigerweise orthonormal sind, sollte es klar sein, daß im allgemeinen Fall das magische Koordinatensystem 255 ein schiefwinkliges sein wird. Es sollte außerdem erwähnt werden, daß sich das magische Koordinatensystem 255, das mit einer starren Substruktur 220 verknüpft ist, stets in einer festen Position und Orientierung bezüglich der letzteren befindet und sich somit zusammen mit der starren Substruktur 220 bewegt, wenn die Substruktur Rotationen und Translationen in dem globalen Koordinatensystem 235 erfährt; dies ist das Ergebnis davon, daß das magische Koordinatensystem 255 konstruktionsgemäß starr an die starre Substruktur 220 angebunden ist. Dasselbe gilt für das schiefwinklige lokale Koordinatensystem 245: aufgrund der Weise, in der es konstruiert ist, befindet sich das magische Koordinatensystem 255 stets in einer festen Position und Orientierung bezüglich des schiefwinkligen lokalen Koordinatensystems 245. Schließlich sollte erwähnt werden, daß es unter Verwendung des magischen Vektorpaares, das mit einer molekularen Substruktur 220 verknüpft ist, eine ganze Skala äquivalenter Weisen zum Aufbau eines magischen Koordinatensystems 255 gibt. Bevor weitergegangen wird, sei in Erinnerung gerufen, daß das schiefwinklige lokale Koordinatensystem 245 unter Verwendung von einigen der Atomplätze der starren molekularen Substruktur 220 erzeugt wurde.

Aus dem obigen ist es klar, daß die Position und die Orientierung des magischen Vektorpaares (und somit des magischen Koordinatensystems) in dem globalen Referenzkoordinatensystem die Position und die Orientierung

der starren Substruktur, z. B. 220, an die das Paar angebunden ist, in dem globalen Referenzkoordinatensystem 235 unambigüen beschreiben.

Sobald ein gegebenes schiefwinkliges lokales Koordinatensystem 245 unter Verwendung von Plätzen von einer molekularen starren Substruktur 220 gebildet ist, kann jeder der Vektoren 238 & 248 des magischen Vektorpaars, das mit der starren Substruktur 220 verknüpft ist, in dem Koordinatensystem 245 dargestellt werden. Diese Darstellung kann entweder explizit oder implizit sein.

In der expliziten Darstellung besitzt der magische Vektor 238 (bzw. 248) eine feste Position und Orientierung in dem ausgewählten schiefwinkligen lokalen Koordinatensystem 245. Diese Position und Orientierung können zum Beispiel in Form eines Translationsvektors T, der den Mittelpunkt des schiefwinkligen lokalen Koordinatensystems 245 mit irgendeinem festen Punkt SP entlang der Achse (Richtung) 217 (bzw. 227) des magischen Vektors 238 (bzw. 248) verbindet, und einer Rotationsmatrix R beschrieben werden. Es sollte klar sein, daß der Punkt SP zum Beispiel einer der Endpunkte D, G (bzw. O, Q) des Vektors 238 (bzw. 248) sein kann. Der Translationsvektor T gibt die Position des Punktes SP in dem schiefwinkligen lokalen Koordinatensystem 245 an, während die Rotationsmatrix R die Orientierung des magischen Vektors 238 (bzw. 248) in dem gleichen schiefwinkligen lokalen Koordinatensystem 245 angibt. Es ist zu erwähnen, daß die Rotationsmatrix durch Auflisten der Längen der Projektionen des magischen Vektors 238 (bzw. 248) auf die Achsen i, j und k des schiefwinkligen lokalen Koordinatensystems 245 äquivalent beschrieben werden kann. Alternativ kann die Rotationsmatrix durch Auflisten der Winkel beschrieben werden, die der magische Vektor 238 (bzw. 248) mit jeder der Achsen i, j und k des schiefwinkligen lokalen Koordinatensystems 245 einschließt. Außerdem können weitere Informationen, z. B. die Identität der drehbaren Bindung 218a (bzw. 218b) — in dem Fall, daß der magische Vektor in Form einer derartigen drehbaren Bindung definiert ist, oder die Größe des Vektors 238 (bzw. 248), in der Darstellung enthalten sein; diese zusätzlichen Informationen können für Verifikationszwecke verwendet werden: zum Beispiel zur Bestimmung der Position einer der zwei anderen Substrukturen 210, 230 bezüglich des schiefwinkligen lokalen Koordinatensystems 245. In dieser Erörterung wurde angenommen, daß allen Bindungen der betrachteten molekularen Struktur eindeutige Bezeichnungen gegeben wurden.

Die Größe, Position und Orientierung der magischen Vektoren 238 & 248 können durch eine geringfügige Modifikation der homogenen 4×4 -Transformationsmatrix, die auf dem Gebiet von Computergraphiken weit verbreitet ist, kompakt in einer Matrixform dargestellt werden. Insbesondere kann die modifizierte Transformationsmatrix, wie das folgende Diagramm anzeigt, durch Verwenden der oben erwähnten 3×3 -Rotationsmatrix R, des 3×1 -Translationsvektors T und der Längen der 3 Projektionen des Vektors 238 (bzw. 248) auf die Achsen i, j und k des schiefwinkligen lokalen Koordinatensystems 245 aufgebaut werden:

$$\begin{bmatrix} \text{ProjektionAufI} \\ \text{ProjektionAufJ} \\ \text{ProjektionAufK} \\ T & 1 \end{bmatrix} R$$

Klarerweise gibt es zwei derartige Matrizen, eine für jedes der zwei Vektorelemente des magischen Vektorpaars.

In der impliziten Darstellung können die Position und Orientierung des magischen Vektorpaars (d. h. der Vektoren 238 und 248) in dem ausgewählten schiefwinkligen lokalen Koordinatensystem 245 durch Auflisten der identifizierenden Bezeichnungen der Atomplätze, die beim Definieren der magischen Vektoren helfen, angegeben werden. Zum Beispiel können in dem Fall, in dem die magischen Vektoren 238 und 248 für die starre Substruktur 220 mit Hilfe der drehbaren Bindungen 218a & 218b definiert werden, die Bezeichnungen D/G und O/Q verwendet werden: die Bezeichnungen sind in der Reihenfolge aufzulisten, welche die Richtung der entsprechenden Vektoren 238 & 248 definiert. Im impliziten Fall werden die Position und Orientierung des magischen Koordinatensystems 255 von den Bezeichnungen der definierenden Atomplätze D/G, O/Q und der Beschreibung der molekularen Struktur 200 erzeugt, wann immer, eine derartige Information über Position und Orientierung notwendig ist. Jegliche weitere Information, wie im Fall der expliziten Darstellung erläutert, kann durch On-line-Berechnung erhalten werden. Daher kann die implizite Darstellung die Speicheranforderungen in dem System 100 vermindern.

Ist entweder die implizite oder die explizite Darstellung der Elemente des magischen Vektorpaars in einem ausgewählten schiefwinkligen lokalen Koordinatensystem 245 gegeben, genügt dies, um die Position und Orientierung der Elemente des magischen Vektorpaars in dem globalen Koordinatensystem 235 zu bestimmen. Wie jedoch oben erläutert wurde, legt eine Bestimmung der Position und Orientierung der Elemente des magischen Vektorpaars in dem globalen Koordinatensystem 235 sofort die Position und Orientierung (d. h. die Platzierung) der entsprechenden starren Substruktur 220 in dem globalen Koordinatensystem 235 fest und spezifiziert diese vollständig.

Wie oben erwähnt, genügt die Definition der Koordinaten (x, y, z) von drei oder mehr Atomen (Plätzen) (z. B. G-P) einer gegebenen starren Substruktur 220 in dem globalen Koordinatensystem 235, um ein schiefwinkliges lokales Koordinatensystem 245 ebenso wie eine globale Position und eine globale Orientierung für die starre Substruktur 220 in dem globalen Koordinatensystem 235 zu definieren. Demzufolge genügt die Definition der Koordinaten (x, y, z) von drei oder mehr Atomen (Plätzen) (z. B. G-P) einer gegebenen starren Substruktur 220 in dem globalen Koordinatensystem 235, um die Position und Orientierung des magischen Vektorpaars in dem

globalen Koordinatensystem 235 zu definieren. Dies wird dadurch erreicht, daß entweder die implizite oder die explizite Darstellung der Vektoren 238 und 248 in dem schiefwinkligen lokalen Koordinatensystem 245 verwendet wird und eine Änderung der Koordinatensysteme durch bekannte Vektortechniken auf das globale Koordinatensystem 235 angewendet wird. In ähnlicher Weise reicht die Definition der Koordinaten (x, y, z) von drei oder mehr Atomen (Plätzen) einer gegebenen starren Substruktur 210 (bzw. 230) in dem globalen Koordinatensystem 235 aus, um die Position und Orientierung des Vektors 238 (bzw. 248) in dem globalen Koordinatensystem 235 zu definieren.

Wie früher angegeben, besitzt die molekulare starre Substruktur 220 ein mit ihr verknüpft magisches Vektorpaar. Entsprechend unserer Beschreibung des magischen Vektorpaars und wenn seine beteiligten Vektoren mit Hilfe von drehbaren Bindungen beschrieben werden, beinhalten die mit den Substrukturen 210 und 230 (die mit der starren Substruktur 220 verbunden sind) verknüpften magischen Vektorpaare die Vektoren 238 beziehungsweise 248. In diesem Moment bleibt die Identität des zweiten Vektors des mit den Substrukturen 210 beziehungsweise 230 verknüpften magischen Paares unspezifiziert. Unten wird geprüft, wie die magischen Vektorpaare definiert sind, wenn eine Substruktur nicht wenigstens zwei drehbare Bindungen aufweist, die von ihr ausgehen. Wenn die Position und Orientierung der Elemente des magischen Vektorpaars für die Substruktur 220 in dem globalen Koordinatensystem 235 bestimmt sind, sind die Position und Orientierung der Substruktur 220 in dem globalen Koordinatensystem 235 vollständig festgelegt, während gleichzeitig die Position und Orientierung der Substrukturen 210 und 230 stark eingeschränkt sind. Tatsächlich erlaubt die Kenntnis der Platzierung der starren Substruktur 220 in dem globalen Koordinatensystem 235 und die Information über die drehbaren Bindungen, die von dieser ausgehen, die Bestimmung der Position und Orientierung der drehbaren Bindungen 218a und 218b in dem globalen Koordinatensystem 235, wie vom Standpunkt der starren Substruktur 220 aus gesehen. In ähnlicher Weise erlaubt das magische Vektorpaar, das mit der starren Substruktur 210 verknüpft ist, die Bestimmung der Platzierung der Substruktur 210 in dem globalen Koordinatensystem 235 und somit die Bestimmung der Position und Orientierung der drehbaren Bindung 218a, wie vom Standpunkt der starren Substruktur 210 aus gesehen. Die zwei Berechnungen müssen jedoch unabhängig davon, von welcher starren Substruktur der Standpunkt zur Bestimmung der Position und Orientierung der drehbaren Bindung 218a in dem globalen Koordinatensystem 235 verwendet wurde, übereinstimmen, wenn die starren Substrukturen Teil einer zutreffenden Anordnung der molekularen Struktur 200 sein sollen. Analoge Kommentare können für die drehbare Bindung 218b und die starren Substrukturen 220 und 230 gemacht werden. Diese Einschränkung ist während der letzten Phase des Vergleichsprozesses 600 (siehe Beschreibung unten) sehr nützlich, da sie es erlaubt, hypothetische Platzierungen für eine gegebene starre Substruktur 220 auszuschließen, wenn diese nicht durch hypothetische Platzierungen von starren Substrukturen 210 und 230, die durch eine drehbare Bindung mit der starren Substruktur 220 verbunden sind, bestätigt werden.

Fig. 2C ist eine Darstellung, die eine zweite molekulare Konformation 250 der molekularen Struktur 200, das globale Koordinatensystem 235, das schiefwinklige lokale Koordinatensystem 245 und die invarianten Vektoren 238 und 248 von Fig. 2A zeigt.

Der Ausdruck Konformation wird verwendet, um irgendeine eines Satzes von möglichen Konfigurationen im dreidimensionalen Raum zu bezeichnen, die eine gegebene molekulare Struktur (200, 250) aufgrund einer inhärenten strukturellen Flexibilität annehmen kann; diese Flexibilität ist typischerweise die Folge von drehbaren und/oder flexiblen Bindungen, die in dem Molekül existieren. Die Analyse hierin konzentriert sich lediglich auf drehbare Bindungen und nimmt an, daß die gezeigte strukturelle Flexibilität das Ergebnis von Drehungen starrer Substrukturen um derartige drehbare Bindungen herum ist. Typischerweise gibt es unendlich viele derartiger Konfigurationen, wobei einige von ihnen energetisch günstiger als andere sind. Außerdem können sterische Betrachtungen den Satz möglicher Konformationen weiter begrenzen.

Wie oben beschrieben, bleiben in jeder Konformation 250 der molekularen Struktur 200 die Position und Orientierung der drehbaren Bindung 218a (bzw. 218b) in bezug auf ein schiefwinkliges lokales Koordinatensystem 245 entweder der Substruktur 210 oder 220 (bzw. 220 oder 230) die gleichen (invariant). In ähnlicher Weise bleiben die Position und Orientierung der drehbaren Bindung 218a (bzw. 218b) in bezug auf ein schiefwinkliges lokales Koordinatensystem 245 entweder der Substruktur 210 oder 220 (bzw. 220 oder 230) die gleichen (invariant), wenn sich die gesamte molekulare Struktur 200 dreht und in dem globalen Koordinatensystem 235 verschiebt. Dies liegt an der Tatsache, daß das schiefwinklige lokale Koordinatensystem 245 und die drehbare Bindung 218a (bzw. 218b) stets in einer festen Position und Orientierung relativ zueinander liegen, ungeachtet jeglicher Rotation 215a (bzw. 215b) der starren Substruktur 220 um die drehbare Bindung 218a (bzw. 218b) herum und jeglicher Translation 295 und/oder Rotation 290 der gesamten molekularen Struktur 200 oder jeder ihrer Konformationen 250.

Um einen Satz beschreibender Indizes für die molekulare Struktur (200, 250) zu erzeugen, müssen Tupel von Atomplätzen (und/oder 'Dummy'-Plätzen, unten beschrieben) gewählt werden. Diese Tupel können zur Bildung des schiefwinkligen lokalen Koordinatensystems 245 verwendet werden. Die Tupel besitzen Tupel-Eigenschaften, die unter anderem geometrische Merkmale, Ordnungs- und Vektorbeziehungen beinhalten können, die durch die Atomplätze, die das Tupel beinhaltet (siehe Beschreibung von Fig. 3), definiert sind.

Außerdem können ein oder mehrere Sätze von Atomplätzen (und/oder 'Dummy'-Plätzen), zum Beispiel der Ring K-P, in der molekularen Struktur (200, 250) als 'Charakteristika' aufweisend identifiziert werden. Diese Charakteristika sind für den Satz von Atomplätzen K-P spezifisch und können umfassen: chemische (z. B. Valenz, Atomgewicht, Atomtyp etc.) und/oder physikalische (z. B. elektrostatische, hydropathische etc.) Eigenschaften des Satzes von Atomplätzen, weitere Eigenschaften etc. Im folgenden werden diese Sätze von Atomplätzen K-P als 'Platzsätze' bezeichnet.

Daher können, wenn ein oder mehr der Atomplätze, die an einem Tupel teilhaben, auch ein Element von einem oder mehr der Platzsätze sind, die Charakteristika der Platzsätze, von denen der Atomplatz in dem ausgewählten

Tupel ein Element ist, auch mit dem Tupel verknüpft sein. Somit können diese Charakteristika dazu verwendet werden, den von dem Tupel abgeleiteten Index zu vergrößern und ihn beschreibender zu machen.

Die Position und Orientierung der magischen Vektoren 238 und 248 sind in jedem der schiefwinkligen lokalen Koordinatensysteme 245, die unter Verwendung von Plätzen von der Substruktur 220 erzeugt werden, dargestellt; die Darstellung der Vektoren 238 und 248, invariant in jedem der schiefwinkligen lokalen Koordinatensysteme 245, ist über eine Datenstruktur (siehe Beschreibung der Fig. 4 und 5 unten) mit dem Index verknüpft, der von dem Tupel abgeleitet ist.

Fig. 3 ist eine Sequenz von Zeichnungen, die zeigen, wie Platzsätze K-O definiert werden (Fig. 3A), wie Dummy-Plätze Du definiert und dann verwendet werden (Fig. 3B) und wie Tupel (typischerweise 335, 345, 355) durch Wählen eines Satzes von einem oder mehr Atomplätzen und/oder Dummy-Plätzen, Du, von der molekularen Struktur 200 definiert werden. Jedes Tupel (335, 345, 355) wird dazu verwendet, ein spezifisches schiefwinkliges lokales Koordinatensystem 245 zu definieren.

Ein Platzsatz ist ein Satz, der ein oder mehr Atomplätze und/oder ein oder mehr Dummy-Plätze der molekularen Struktur 200 beinhaltet. Ein Beispiel für einen Platzsatz kann eine allgemein auftretende Struktur (z. B. ein Phenylring oder der Ring K-O) in einer Datenbank, D, aus molekularen Strukturen 200 sein. Manchmal ist es nützlich, eine derartige Struktur durch einen einzelnen Dummy-Platz Du zu ersetzen. Ein alternativer Weg zur Definition eines Platzsatzes besteht darin, Atome auszuwählen, die einen gemeinsamen Satz von Charakteristika und/oder Eigenschaften teilen. Zum Beispiel kann man einen Platzsatz durch Sammeln aller der Atomplätze erzeugen, die an einem aromatischen Ring teilhaben. Ein weiterer Platzsatz kann durch Sammeln all jener Plätze erzeugt werden, die als Wasserstoff-Donatoren (beziehungsweise Akzeptoren) wirken. Diese Platzsätze können auch durch eine Dummy-Einheit ersetzt werden, wobei in diesem Fall die Dummy-Einheit alle die Charakteristika des Platzsatzes übernimmt, der ersetzt wird. Zum Beispiel ist in Fig. 3A der Platzsatz K-O an den Atomplatz P gebunden. Außerdem existiert eine drehbare Bindung, welche die Plätze O und Q verbindet. Wenn der Platzsatz K-O durch die Dummy-Einheit Du (Fig. 3B) ersetzt wird, ist es die Dummy-Einheit anstelle des Platzsatzes K-O, die nun an den Atomplatz P gebunden ist. In einer analogen Weise wird nun die drehbare Bindung OQ durch eine Bindung ersetzt, welche die Dummy-Einheit mit dem Platz Q in der Substruktur 230 verbindet. Außerdem übernimmt die Dummy-Einheit Du, wenn der Platzsatz K-O Charakteristika besitzt (z. B. hydropathisches, bestimmtes elektrostatisches Verhalten etc.) auch diese Charakteristika.

Ein Tupel ist ein Satz von einem oder mehr Atomplätzen und/oder einem oder mehr Dummy-Plätzen. Tupel, die lediglich einen Atom-(oder Dummy-)Platz beinhalten, sind bei der Beschreibung von Translationen 295 einer starren Struktur nützlich. In derartigen Fällen kann eine Bestimmung von Information über die Rotation 290 eine zusätzliche berechnungsmäßige Belastung nach sich ziehen. Des weiteren können Informationen zur Erzeugung von Indizes auf die Charakteristika des einzelnen Atom-(oder Dummy-)Platzes in dem Tupel beschränkt sein. In ähnlicher Weise sind Tupel, die lediglich zwei Atom-(oder Dummy-)Plätze beinhalten, bei der Beschreibung von Translationen 295 einer starren Struktur nützlich, und sie können auch die Rotation 290 auf zwei Freiheitsgrade beschränken, spezifizieren die Rotation jedoch nicht vollständig: eine Bestimmung von Information über die Rotation 290 zieht eine zusätzliche berechnungsmäßige Belastung nach sich. In diesem Fall können Informationen zur Erzeugung von Indizes auf die Charakteristika der zwei Atom-(oder Dummy-)Plätze in dem Tupel beschränkt sein.

Bei der bevorzugten Ausführungsform werden Tupel unter Verwendung von drei oder mehr Atom-(und/oder Dummy-)Plätzen definiert. Bei einer bevorzugteren Ausführungsform sind wenigstens drei Atom-(und/oder Dummy-)Plätze das Tupels nicht kollinear. Die Tupel werden dazu verwendet, ein schiefwinkliges lokales Koordinatensystem 245 (wie oben beschrieben) und einen Index zu definieren. Wenn das Tupel vier Atom-(und/oder Dummy-)Plätze beinhaltet, von denen jeweils drei nicht kollinear sind, dann kann das Kreuzprodukt $i \times j$ (oben beschrieben) durch den Vektor ersetzt werden, der den Ursprung des schiefwinkligen lokalen Koordinatensystems 245 mit dem vierten Platz verbindet.

Es ist zu erwähnen, daß vier oder mehr Atom-(und/oder Dummy-)Plätze verwendet werden können. In diesem Fall können jeweils drei nicht kollineare Atom-(und/oder Dummy-)Plätze ausgewählt werden, um das schiefwinklige lokale Koordinatensystem 245 zu bestimmen, während die restlichen Atom-(und/oder Dummy-)Plätze dazu verwendet werden können, die während der Vergleichsstufe des Verfahrens erzeugten Hypothesen weiter einzuschränken. Siehe Beschreibung von Fig. 6 unten.

Man beachte, daß keine, einige oder alle dieser Eigenschaften und Charakteristika dazu verwendet werden können, eine Zahl (einen Index) zu bilden, der die erzeugten Dreiecke, die den Tupeln entsprechen, eindeutig beschreibt.

Die magischen Vektoren 238 und 248 sind starr in dem schiefwinkligen lokalen Koordinatensystem 245 angeordnet, das unter Verwendung von Plätzen von der Substruktur 220 gebildet wurde, wie oben beschrieben. Die magischen Vektoren 238 und 248 werden dann (implizit oder explizit, wie oben beschrieben) in dem schiefwinkligen lokalen Koordinatensystem 245 des gebildeten Dreiecks 300 dargestellt.

Tupel werden während der Durchführung von zwei Prozessen, die in dieser Erfindung enthalten sind, gebildet: einem Referenzspeicherprozeß (siehe Fig. 5) und einem Vergleichsprozeß (siehe Fig. 6). In dem Referenzspeicherprozeß werden Tupel durch Auswählen von Atom-(und/oder Dummy-)Plätzen aus der molekularen Struktur (200, 250) gebildet. Während des Referenzspeicherprozesses werden die Tupel durch Auswählen aus einem Satz von Atom-(und/oder Dummy-)Plätzen, Referenz-Tupel-Auswahlsatz genannt, gebildet. Der Referenz-Tupel-Auswahlsatz beinhaltet alle Atomplätze in einer starren Substruktur (210, 220, 230), alle Dummy-Plätze, die mit einer starren Substruktur (210, 220, 230) verknüpft sind, und die Atom-(und/oder Dummy-)Plätze, die Endpunkte jeglicher drehbarer Bindungen 218a und 218b sind, die an die gegebene starre Substruktur (210, 220, 230) angebunden, jedoch nicht in der Substruktur (210, 220, 230) enthalten sind. Der Referenz-Tupel-Auswahlsatz beinhaltet diese Atom- (und/oder Dummy-)Plätze, da Tupel, die einen oder mehr dieser Plätze enthalten,

invariant bleiben, ungeachtet der Rotation 215 um irgendwelche drehbaren Bindungen 218a und/oder 218b herum. Dies liegt daran, wie oben erläutert, daß die Position und Orientierung der drehbaren Bindung 218a (bzw. 218b) in bezug auf das schiefwinklige lokale Koordinatensystem 245, welches von dem Tupel definiert wird, gleich (invariant) bleiben, wenn die starre Substruktur 220 bezüglich entweder der starren Substruktur 210 oder der drehbaren Bindungen 218a und 218b in dem schiefwinkligen lokalen Koordinatensystem 245, welches von dem Tupel definiert wird, invariant sind, ungeachtet jeglicher Rotation 290 und Translation 295 der molekularen Struktur (200, 250). In einer bevorzugten Ausführungsform können die Tupel aus einem geeigneten Subsat des soeben definierten Referenz-Tupel-Auswahlsatzes gewählt werden.

Während des Vergleichstestprozesses werden die Tupel durch Auswählen aus einem Satz von Atom-(und/oder Dummy-)Plätzen gebildet, welcher der 'Vergleichstupel-Auswahlsatz' genannt wird. Anders als bei dem Referenz-Tupel-Auswahlsatz kann der Vergleichstupel-Auswahlsatz alle Atom-(und/oder Dummy-)Plätze der gesamten molekularen Struktur (200, 250) beinhalten. Bei einer alternativen bevorzugten Ausführungsform können Sätze aller dieser Plätze verwendet werden, um den Vergleichstupel-Auswahlsatz zu erzeugen. Bei einer bevorzugten Ausführungsform beinhalten die Atomplätze sowohl in dem Referenz-Tupel-Auswahl- als auch dem Vergleichstupel-Auswahlsatz keine Atom- (und/oder Dummy-) Plätze, die zu weit voneinander entfernt sind (z. B. mehr als 10 Angström entfernt).

Die Erfindung erzeugt eine Mehrzahl von Tupeln sowohl in dem Referenzspeicher- 500 als auch dem Vergleichsprozess 600. Bei einer bevorzugten Ausführungsform werden so viele Tupel wie möglich durch Verwenden des Referenz-Tupel-Auswahlsatzes (oder des Vergleichstupel-Auswahlsatzes) erzeugt. In einer weiteren bevorzugten Ausführungsform werden alle möglichen Tupel erzeugt, die durch diese Tupelauswahlsätze impliziert sind. Bei einer weiteren bevorzugten Ausführungsform werden alle möglichen Tupel mit Ausnahme redundanter Permutationen der Tupelmitglieder erzeugt, die durch diese Tupelauswahlsätze impliziert sind.

Fig. 3C ist eine Darstellung der Substruktur 220 mit angeordneten drehbaren Bindungen 218a und 218b und der Ringstruktur K-O, die durch den Dummy-Platz Du dargestellt ist. Ein Tupel 335 wird durch Auswahl von drei Atomplätzen H, I, J aus dem Referenz-Tupel-Auswahlsatz gebildet, der den Satz von Atomplätzen D, G, H, I, J, P und den Dummy-Platz Du beinhaltet. Das Tupel 335 definiert ein Dreieck 336 mit Eigenschaften, die folgendes umfassen: geometrische Merkmale (z. B. die Längen der drei Seiten des Dreiecks 336, die Winkel des Dreiecks 336, den Umfang des Dreiecks 336 etc.), Ordnungsinformation (die durch Konvention aus der Reihenfolge, in der die Plätze ausgewählt werden, impliziert ist), Vektorinformation etc. Zum Beispiel ist, wenn die Atomplätze in der Reihenfolge H, I, J ausgewählt werden, der Vektor i (siehe Erörterung oben) als $i = H \rightarrow I$ definiert, und der Vektor j (siehe Erörterung oben) ist als $j = H \rightarrow J$ definiert; diese Konvention wird durch den ganzen beschriebenen Prozeß hindurch konsistent verwendet. Andere Konventionen sind möglich. Alternativ ist, wenn die Atomplätze in der Reihenfolge I, H, J ausgewählt werden, der Vektor i (siehe Erörterung oben) als $i = I \rightarrow H$ definiert, und der Vektor j (siehe Erörterung oben) ist als $j = I \rightarrow J$ definiert. In beiden Fällen ist der Vektor k als $k = i \times j$ definiert, wie oben beschrieben, und die Vektoren i, j, k definieren das schiefwinklige lokale Koordinatensystem 245, das mit dem Tupel 335 verknüpft ist.

Wenn eine gegebene Anzahl, z. B. 3, von Atomplätzen aus dem Referenz-Tupel-Auswahlsatz ausgewählt wird, ist mehr als eine Tupelreihenfolge möglich. Mit anderen Worten können die ausgewählten Plätze, die das Tupel bilden, permutiert werden, um weitere Tupel zu bilden. Zum Beispiel können die ausgewählten Atomplätze H, I, J Tupel 335 wie folgt bilden: H-I-J, H-J-I, I-H-J, I-J-H, J-I-H und J-H-I. Im allgemeinen ist die Anzahl von geordneten Tupeln, die durch Auswahl von k Plätzen aus einem Referenz-Tupel-Auswahlsatz erzeugt werden können, der l Plätze enthält, durch $l/(l-k)!$ gegeben.

Tupel 335, die Permutationen voneinander sind, definieren jedoch die gleichen geometrischen Eigenschaften, z. B. die Längen der Seiten des Dreiecks 336 etc. Daher sind in einigen bevorzugten Ausführungsformen redundante Permutationen einer gegebenen Anzahl von Plätzen, die ein Tupel bilden, nicht notwendig. Dies liegt daran, daß alle Permutationen eines gegebenen Tupels 335 den gleichen Satz von Atomplätzen beinhalten und daher die gleichen geometrischen Merkmale und Vektorinformationen tragen.

Ordnungsinformationen können auch erhalten werden, wenn eine Ordnungskonvention auferlegt wird: Alle Permutationen eines gegebenen Tupels 335 können aus einer einzigen normierten Form des Tupels 335 unter Verwendung der Ordnungskonvention erzeugt werden. Daher genügt es, lediglich ordnungsfreie Kombinationen von Atomplätzen von der molekularen Struktur (200, 250) zu berücksichtigen; die Anzahl möglicher (ordnungsfreier) Kombinationen, die durch Wählen von k Plätzen aus einem Referenz-Tupel-Auswahlsatz, der l Plätze enthält, erzeugt werden kann, ist durch $l/(k!(l-k)!)!$ gegeben, was um einen Faktor k! kleiner als die Anzahl geordneter Tupel ist. Demgemäß sind die Speicheranforderungen um den gleichen Faktor reduziert, bei einer minimalen Zunahme der berechnungsmäßigen Kosten, die zur Ausführung der notwendigen Verwaltungsoperationen notwendig sind.

Der Prozeß des Auswählens eines einzelnen repräsentativen ordnungsfreien Tupels (= einer Kombination) wird 'Normierung' genannt. Eine Normierung beinhaltet eine Bestimmung einer eindeutigen Ordnung, wenn ein Satz von Atomplätzen gegeben ist; die auferlegte Ordnung ist unabhängig von der Reihenfolge, in der die Atomplätze gegeben sind. Dies wird durch Auferlegen einer Ordnungskonvention bewerkstelligt, um ein einzelnes repräsentatives 'normiertes' Tupel bei einem gegebenen Satz von Atomplätzen auszuwählen. Die Atomplätze werden gemäß einer bevorzugten Ordnungskonvention dadurch geordnet, daß zuerst die aktuellen Längen der Seiten der Form bestimmt werden, die durch Verbinden der ausgewählten Plätze erzeugt wird. Es sind auch andere Ordnungskonventionen möglich. Der erste und zweite Platz in der Reihenfolge sind jene Plätze, die am weitesten entfernt sind und die längstmögliche Seite eines Polygons bilden, das jeden der ausgewählten Plätze als einen Vertex besitzt. Der dritte Platz in der Reihenfolge ist der Platz, der am weitesten von jedem der ersten beiden Plätze entfernt ist und die nächstlängste Seite des Polygons bildet. Der zweite Platz in der Reihenfolge

wird dann der Platz an dem Vertex, an dem sich die zwei zuvor gebildeten Seiten schneiden. Der erste Platz in der Reihenfolge wird dann der andere Platz auf der längstmöglichen Seite. Das Ordnen wird fortgesetzt durch Wählen des vierten Platzes als jenem verbliebenen Atomplatz, der den weitesten Abstand von dem dritten Platz hat, des fünften Platzes als dem verbliebenen Atomplatz, der am weitesten von dem vierten Platz entfernt ist und so weiter, bis alle Plätze des Tupels 335 geordnet sind.

Zum Beispiel kann unter Verwendung eines Tupels 335 aus drei Atomplätzen H, I und J ein Dreieck 336 gebildet werden, dessen Seiten gemäß der oben beschriebenen Ordnungskonvention geordnet sind. Um dies zu bewerkstelligen, bildet der längste Abstand I-J zwischen jeweils zwei der Plätze H, I, J die längste Seite des Dreiecks 336. Die zweite Seite wird durch den größten Abstand entweder von I oder J zu dem verbliebenen Platz H bestimmt; in diesem Fall ist dieser Abstand I-H. Als Folge ist der zweite Platz in der Reihenfolge I, da er sich am Vertex befindet, den sich I-J und I-H teilen; der erste Platz in der Reihenfolge ist J, welches den anderen Vertex auf der längsten Seite darstellt; und der dritte Platz in der Reihenfolge ist der als einziges verbliebene Platz, H.

Es ist zu erwähnen, daß Verbesserungen für die Konvention notwendig sind, um jegliche existierenden Symmetrien in dem Polygon zu brechen. Zum Beispiel kann, wenn die Seiten I-H und I-J die gleiche Länge haben, das Ordnen nicht auf dem Abstand allein basieren, sondern es sind weitere Kriterien zu verwenden. Diese Kriterien können auf anderen Eigenschaften des Tupels basieren, wie Ordnungszahlen der Atome an den Plätzen, chemische Eigenschaften etc. Zum Beispiel ist in dem Fall, in dem I-H und I-J die gleiche Länge haben, der Atomplatz I der zweite in der Reihenfolge, da er den Vertex darstellt, den sich die längste und die zweitlängste Seite (gleiche Seiten) teilen. Die Reihenfolge der Atomplätze J und H ist jedoch zweideutig und kann zum Beispiel durch Auswählen jenes Platzes von J und H mit der höchsten Ordnungszahl als dem ersten Platz in der Reihenfolge gelöst werden. Ähnliche Überlegungen können in dem Fall verwendet werden, in dem das Dreieck 336 gleichseitig ist.

Sobald das Tupel 335 normiert ist, wie oben beschrieben, wird ein eindeutiger Index gebildet, der das Tupel 335 repräsentiert. Dieser Index kann unter Verwendung einer beliebigen Anzahl von geometrischen Merkmalen, Eigenschaften der Plätze des Tupels, chemischen und/oder physikalischen Informationen über das Tupel oder die Atomplätze des Tupels etc. erzeugt werden. Man nehme zum Beispiel an, daß der Atomplatz J ein doppelt gebundenes Stickstoffatom, I ein einfach gebundenes Kohlenstoffatom und der Atomplatz H ein doppelt gebundenes Kohlenstoffatom sind. Des weiteren ist zu beachten, daß das Dreieck 336 eine längste Seite mit der Länge l_1 , eine zweitlängste Seite mit der Länge l_2 und eine dritte Seite mit der Länge l_3 besitzt. In ähnlicher Weise besitzt das Dreieck 336 Winkel θ_1 , θ_2 beziehungsweise θ_3 , die den geordneten Atomplätzen J, I und H entsprechen. Bei gegebener derartiger Information kann ein eindeutiger Index unter Verwendung keiner oder mehr der Seiten l_1 bis l_3 , keinem oder mehr der Winkel θ_1 bis θ_3 , keiner oder mehr der Bindungsartbezeichnungen (einfach gebunden, doppelt gebunden etc.), keinem oder mehr der chemischen Typen (Stickstoff, Kohlenstoff etc.) und/oder keinem oder mehr der physikalischen Eigenschaften (Atomgewicht der Atome an den Plätzen, Elektronegativität etc.) etc. erzeugt werden, der dieses Tupel J-I-H beschreibt. Bei einer bevorzugten Ausführungsform wird der Index durch Verwenden der Längen l_1 , l_2 , des Winkels θ_2 und des Atomtyps des Atoms am zweiten Platz in der Ordnung erzeugt. Bei anderen Ausführungsformen kann es erwünscht sein, Indizes dann und nur dann zu erzeugen, wenn die Längen l_1 und/oder l_2 einen bestimmten Schwellwert überschreiten und/oder der Winkel θ_2 einen bestimmten Schwellwert überschreitet; typische Schwellwerte können 1 Angström für die Längenabmessung und 10 Grad für die Winkelabmessung sein. Schließlich kann es hin und wieder wünschenswert sein, den Index durch Verwenden der Längen l_1 , l_2 und des größten Winkels in dem durch das Tupel gebildeten Dreieck 336 zu erzeugen.

In Anbetracht der obigen Erörterung werden diese Tupel dadurch erzeugt, daß der Referenz-Tupel-Auswahlsatz während des Referenzspeicherprozesses 500 und der Vergleichstupel-Auswahlsatz während des Vergleichsprozesses 600 verwendet werden. Bei einer bevorzugten Ausführungsform wird jede mögliche Kombination von beteiligten Plätzen in entweder dem Referenz-Tupel-Auswahlsatz oder dem Vergleichstupel-Auswahlsatz erzeugt. Bei alternativen Ausführungsformen können weniger Tupel erzeugt werden. Zum Beispiel wird in Fig. 3D ein Tupel 345 durch Atomplätze I, J und den Dummy-Platz Du gebildet. Dieses Tupel wird normiert, wie oben erläutert, und der entsprechende eindeutige Index wird erzeugt. In einer ähnlichen Weise wird jedes andere mögliche Tupel, typischerweise 355 (DGI) in Fig. 3E, gebildet, normiert, und ein Index wird erzeugt. Man beachte, daß jeder dieser Indizes für das zugehörige Tupel eindeutig und invariant bezüglich Translation 295 und Rotationen 290 der molekularen Struktur (200, 250) und jeglichen Rotationen 215a von jeder molekularen Substruktur (210, 220) um irgendeine drehbare Bindung 218a herum ist.

Außerdem wird für jedes gebildete Tupel (335, 345, 355) in der oben beschriebenen Weise ein mit dem Tupel verknüpft schiefwinkliges lokales Koordinatensystem 245 abgeleitet. Mit den magischen Vektoren 238 & 248 ist Vektorinformation verknüpft und ist in jedem der schiefwinkligen lokalen Koordinatensysteme 245 dargestellt. Daher sind die Vektorinformation, die Identität der molekularen Struktur 200, die Identitäten der molekularen Substrukturen (210, 220, 230), die Identitäten der drehbaren Bindungen 218a und 218b, der Index 414, das Tupel 335 und das schiefwinklige lokale Koordinatensystem 245 alle miteinander verknüpft.

Vektorinformation ist Information über gegebene magische Vektoren 238 & 248 und beinhaltet die Darstellungen der magischen Vektoren in dem schiefwinkligen lokalen Koordinatensystem 245. Bei einer bevorzugten Ausführungsform ist diese Vektorinformation die explizite und/oder implizite Darstellung der magischen Vektoren, wie oben beschrieben (Fig. 2A).

Es ist zu erwähnen, daß ein ausgewähltes Tupel 335 und das zugehörige gebildete Dreieck 336 auch in einer anderen molekularen Struktur als 200 auftauchen kann. Dies erfordert die Verbesserung der Vektorinformation mit der Berücksichtigung der oben beschriebenen Identität der molekularen Struktur; diese Berücksichtigung erlaubt die Identifizierung der einzelnen molekularen Struktur 200, der die Vektorinformation entspricht.

Fig. 4 ist ein Blockdiagramm einer Datenstruktur 400, die einen Index 414, der zu einem Tupel (typischerweise 335, 345, 355) gehört, mit Information über die Identitäten der Atomplätze, die an dem Tupel teilhaben, und Information verknüpft, die zu den Darstellungen 238A und 248A von magischen Vektoren 238 und 248 in dem schiefwinkligen lokalen Koordinatensystem 245 des das Tupel erzeugenden, Index 414 gehört. Man beachte, daß das Tupel, das mit dem Index 414 verknüpft ist, mehr als einmal in einer molekularen Struktur (200, 250) oder in mehr als einer molekularen Struktur (200, 250) in einer Datenbank D auftauchen kann, die eine Mehrzahl von molekularen Strukturen (200, 250) enthält. Als eine Folge davon gibt es im allgemeinen mehr als einen Eintrag 412 von Vektorinformation in einen Datensatz 425 der Datenstruktur 400. Demzufolge beinhaltet jeder derartige Eintrag, typischerweise 412, von Vektorinformation Identifikationsinformationen für jede der molekularen Strukturen 421A bis 421N, in denen der das Tupel erzeugende Index 414 auftaucht. Ein Datensatz 425 enthält außerdem das Koordinatensystemtupelfeld, das alle die Informationen beinhaltet, die zu dem Index 410 gehören, das Koordinatensystemtupel, das es erzeugte, und potentiell weitere Informationen.

Wie oben beschrieben, wird der eindeutige Index 414 erzeugt, der das Tupel 335 darstellt. Dieser Index 414 kann unter Verwendung einer beliebigen Anzahl geometrischer Merkmale, Eigenschaften der Plätze des Tupels, chemischer und/oder physikalischer Informationen über das Tupel oder die Atomplätze des Tupels etc. erzeugt werden. Außerdem kann dieser Index einem Offset in einem eindimensionalen linearen Datenfeld wie 400 zugeordnet werden, indem übliche Offset-Berechnungsverfahren (z. B. 'Schritt'-Berechnung) verwendet werden. Zum Beispiel wird unter Verwendung von I1, I2, Ø2 und dem SYBYL-Atomtyp des zweiten Atomplatzes in der (normierten) Reihenfolge zur Erzeugung eines Index der berechnete Offset (d. h. die Stelle in der Datenstruktur 400) wie folgt bestimmt:

1. Man quantisiert den Wert V_{Ai} jeder Eigenschaft A_i ($i=1, 2, 3, 4, \dots$), indem man den ganzzahligen Wert des Ausdrucks:

$$\frac{V_{Ai} - \min(A_i)}{\max(A_i) - \min(A_i)} \times (\text{STEPS}(A_i) - 1)$$

nimmt, wobei $\min(A_i)$ der minimale Wert ist, der für die Eigenschaft A_i erlaubt ist, $\max(A_i)$ der maximale Wert ist, der für die Eigenschaft A_i erlaubt ist, $\text{STEPS}(A_i)$ die von den Daten abhängige Anzahl von Quantisierungsschritten ist, in die das Intervall $[\min(A_i), \max(A_i)]$ unterteilt ist (diese Anzahl von Schritten wird vor dem Anwenden des Verfahrens festgelegt und fixiert), und i den Satz von Eigenschaften durchläuft, die zur Erzeugung des Index 414 verwendet werden. Beispiel: Wenn die Längeneigenschaft I1 den Wert 1,3 Angström besitzt und unter der Annahme, daß der Bereich möglicher Werte, der sich von 0,9 Angström bis 4,5 Angström erstreckt, in 64 Quantisierungsschritte unterteilt wurde, ist der abgeleitete quantisierte Wert für I1:

$$\left\lfloor \frac{1,3 - 0,9}{4,5 - 0,9} \times (64 - 1) \right\rfloor = 6.$$

In einer ähnlichen Weise wird der quantisierte Wert jeder Eigenschaft A_i bestimmt. Es ist zu erwähnen, daß für Eigenschaften A_i , die inhärent Werte aus einem endlichen Satz von ganzen Zahlen annehmen können (z. B. die 41 SYBYL-Atomarten), die Variable $\text{STEPS}(A_i)$ auf die Kardinalität dieses Satzes von ganzen Zahlen reduziert werden kann.

2. Man nimmt die quantisierten Werte A_i und rechnet unter Verwendung eines 'Schritt'-Berechnungsverfahrens den Offset in die lineare Anordnung 400 hinein. In diesem speziellen Beispiel liegen die folgenden Entsprechungen vor: $A_1 \leftrightarrow I1$, $A_2 \leftrightarrow I2$, $A_3 \leftrightarrow \delta$, $A_4 \leftrightarrow \text{SYBYL-Atomart}$. Die Berechnung des Offsets ergibt:

$$\text{offset} = \left\lfloor \begin{aligned} &A_1 \cdot (\text{STEPS}(A_2) \cdot \text{STEPS}(A_3) \cdot \text{STEPS}(A_4)) \\ &\quad + A_2 \cdot (\text{STEPS}(A_3) \cdot \text{STEPS}(A_4)) \\ &\quad \quad + A_3 \cdot \text{STEPS}(A_4) \\ &\quad \quad \quad + A_4 \end{aligned} \right\rfloor$$

Die Struktur 400 wird durch Prozesse 500 und 600 verwendet, wie unten beschrieben. Die Erörterung nahm bislang implizit an, daß (i) jede molekulare starre Substruktur (210, 220, 230) wenigstens zwei an diese angebundene drehbare Bindungen 218a und 218b besitzt; (ii) jede molekulare starre Substruktur (210, 220, 230) ein mit ihr verknüpftes magisches Vektorpaar aufweist; und (iii) daß die Elemente des magischen Vektorpaars manchmal mit Hilfe von drehbaren Bindungen definiert sind.

Soweit es die Annahme (i) betrifft, kann dies klarerweise nicht immer der Fall sein. Tatsächlich sind die folgenden Situationen ebenfalls möglich:

- (a) starre Substrukturen ohne drehbare Bindungen; mit anderen Worten, einige molekulare Strukturen in der Datenbank D können starr sein und keine drehbaren Bindungen enthalten;
 (b) starre Substrukturen mit einer drehbaren Bindungen, die von der Substruktur ausgeht; dies ist zum Beispiel der Fall der starren Substrukturen 210 und 230 in Fig. 2; und
 (c) starre Substrukturen mit mehr als zwei drehbaren Bindungen, die von der Substruktur ausgehen.

5

Die Tatsache, daß molekulare Strukturen mit einer oder mehr der obigen drei Charakteristika ebenfalls in der Datenbank D vorhanden sind, erfordert eine leichte Modifikation der oben beschriebenen Prozedur zur Index-erzeugung.

Im Hinblick auf die Annahme (ii) wird als nächstes erläutert, wie das magische Vektorpaar bestimmt wird. In dem Fall von starren molekularen Strukturen, die keine drehbaren Bindungen enthalten (Fall (a) oben), kann das magische Vektorpaar nicht mit Hilfe der drehbaren Bindungen erzeugt werden. Statt dessen kann das magische Vektorpaar in einer bevorzugten Ausführungsform leicht durch Identifizieren von zwei Paaren von Atomplätzen definiert werden: ein derartiges Paar von Plätzen kann zum Beispiel durch 1) das Paar von Atomplätzen, die in der betrachteten molekularen Struktur am weitesten voneinander entfernt sind, und 2) das Paar von Atomplätzen gebildet werden, die so weit wie möglich voneinander entfernt liegen und eine Richtung festlegen, die so orthogonal wie möglich zu der durch das erste Paar definierten Richtung ist. Außerdem sind Modifikationen dieser Prozedur möglich: Das Hauptziel hier ist die Erzeugung eines nicht entarteten magischen Koordinatensystems. Die Linie, welche die zwei Plätze des ersten Paares verbindet, entspricht der Achse des Vektors 238. Dies kann als äquivalent zum Vorliegen einer 'fiktiven drehbaren' Bindung 218a angesehen werden, welche die starre Substruktur 210 mit sich selbst verbindet. Die Richtung des Vektors 238 kann jedoch nicht länger durch Verwenden der Bezeichnungen der fraglichen starren Substrukturen bestimmt werden: eine bevorzugte Ausführungsform führt eine Modifikation ein, gemäß der die Richtung von 238 durch Verwenden der Anzahl der Atomplätze bestimmt wird, welche die (fiktive drehbare) Bindung verbindet: Es wird angenommen, daß die Konvention für die Richtung konsistent für alle der einen oder mehr analysierten molekularen Strukturen vom Atomplatz mit der kleineren (größeren) Zahl zu dem Atomplatz mit der größeren (kleineren) Zahl verläuft. In einer ähnlichen Weise entspricht die Linie, welche die zwei Plätze des zweiten Paares verbindet, der Achse des Vektors 248; die Richtung von 248 wird durch die gleiche Konvention bestimmt, die zur Bestimmung der Richtung des Vektors 238 verwendet wurde.

Analog kann in dem Fall von starren molekularen -Substrukturen mit einer drehbaren Bindung, die von der Substruktur ausgeht (Fall (b) oben), einer der magischen Vektoren mit Hilfe der vorhandenen drehbaren Bindung definiert werden, während der zweite magische Vektor die Erzeugung einer 'fiktiven drehbaren' Bindung, wie oben erläutert, erfordert. In einer bevorzugten Ausführungsform kann dies leicht durch Identifizieren eines Paares von Atomplätzen in der betrachteten starren molekularen Substruktur erreicht werden, mit der offensichtlichen Einschränkung, daß das durch das magische Vektorpaar erzeugte magische Koordinatensystem nicht entartet ist. Die Richtung der (fiktiven drehbaren) Bindung und somit des zweiten magischen Vektors wird, wie bereits oben erläutert, durch Konvention bestimmt. In einer alternativen Ausführungsform können beide magische Vektoren in Form von fiktiven drehbaren Bindungen definiert werden.

Im Fall von starren molekularen Substrukturen mit mehr als zwei drehbaren Bindungen, die von der Substruktur ausgehen (Fall (c) oben), gibt es eine Wahl dahingehend, wie das magische Koordinatensystem (äquivalent: das magische Vektorpaar) definiert wird. Zum Beispiel können zwei der drehbaren Bindungen bei der Definition des magischen Vektorpaares verwendet werden; klarerweise genügt jedes beliebige Paar drehbarer Bindungen, das zu einem nicht entarteten magischen Koordinatensystem führt. Alternativ kann ein magischer Vektor mit Hilfe einer der drehbaren Bindungen definiert werden, während der zweite mit Hilfe einer 'fiktiven drehbaren' Bindung definiert werden kann. Oder es können beide magischen Vektoren in Form von 'fiktiven drehbaren' Bindungen definiert werden.

Hin und wieder und für jene starren Substrukturen 220, die mehr als zwei drehbare Bindungen aufweisen, die von der Substruktur ausgehen, kann es wünschenswert sein, eine gewisse Redundanz einzuführen und mehr als zwei, mit der starren Substruktur 220 zu verknüpfende, magische Vektoren zu definieren; das magische Koordinatensystem kann dann in Form von beliebigen zwei nicht kollinearen Vektoren aus dem Satz magischer Vektoren definiert werden. In bestimmten Ausführungsformen können die impliziten oder expliziten Darstellungen all dieser magischen Vektoren in einen Eintrag 412 der Datenstruktur 400 aufgenommen werden.

Was die Annahme (iii) betrifft, ist es aus der Analyse der vorherigen paar Absätze klar, wie die Verwendung der drehbaren Bindungen bei der Definition magischer Vektoren ausgedehnt werden kann. Von einem berechnungsmäßigen Standpunkt aus erleichtert die Definition eines magischen Vektors mit Hilfe einer drehbaren Bindung des weiteren die Erhärtung konsistenter Ergebnisse während der Stufe des Vergleichens, wobei Konformationen in Übereinstimmung mit so vielen erzeugten Antworten (d. h. hypothetische Plazierungen für die jeweilige starre Substruktur) wie möglich bestimmt werden (für eine Erläuterung siehe unten). Gelegentlich kann es wünschenswert sein, einen Eintrag 412 der Datenstruktur 400 um die Darstellung von einer oder mehreren der drehbaren Bindungen, die von einer starren Substruktur ausgehen, in einem schiefwinkligen lokalen Koordinatensystem zusätzlich zu der Darstellung der magischen Vektoren zu vergrößern; mit anderen Worten enthält ein Eintrag der Datenstruktur 400 die Darstellung in einem schiefwinkligen lokalen Koordinatensystem von zwei oder mehr Vektoren, die starr an die starre Substruktur angebunden sind, zu der das schiefwinklige lokale Koordinatensystem gezeichnet ist. Siehe auch die Beschreibung der Verwendung der Abstimmtable unter.

Als Schlußfolgerung ist zu erwähnen, daß es, wenn fiktive magische Vektoren erzeugt werden, Situationen geben kann, in denen es notwendig ist, Symmetrien zu brechen. Zum Beispiel: molekulare Strukturen, die zwei Paare von Atomplätzen enthalten, deren Paarelemente sich in gleichem Abstand voneinander befinden. Um

diesen Problempunkt zu lösen, kann eine Ausführungsform zum Beispiel das Paar, das den Atomplatz mit der niedrigsten Nummer enthält, behalten und den anderen ausscheiden.

Mit den beschriebenen Modifikationen kann nun jede molekulare starre Substruktur 210, 220, 230 einer molekularen Struktur 200, 250 mit einem Paar magischer Vektoren 238 und 248 verknüpft werden, die wie üblich in jedem schiefwinkligen lokalen Koordinatensystem 245 ausgedrückt werden können, die aus dem Referenz-tupel-Auswahlsatz erzeugt werden können.

Der Prozeß 500 analysiert einen Satz aus einer oder mehreren molekularen Strukturen (200, 250) in einer Datenbank D, die eine Mehrzahl von molekularen Strukturen (200, 250) enthält, indem für eine Mehrzahl von Indizes 414 keine oder mehr molekulare Strukturen (200, 250) und/oder Substrukturen (210, 220, 230) bestimmt werden, die Tupel enthalten, die bezüglich der Eigenschaften A_i ähnlich sind, die dazu verwendet werden, den Index 414 zu bestimmen, für den jede dieser molekularen Strukturen (200, 250) und/oder Substrukturen (210, 220, 230) einen Eintrag 412 erzeugt:

- (a) Identifizieren (421A bis 421N) einer der molekularen Strukturen (200, 250), die mit einem gegebenen magischen Vektorpaar 238 und 248 verknüpft ist;
- (b) Identifizieren des Koordinatensystemtupels, das den Index 414 erzeugte;
- (c) Identifizieren (422A bis 422N) der starren molekularen Substruktur (210, 220, 230), aus der das Tupel an der Datenfeldstelle der Struktur 400 entnommen wurde, die zu dem Index 414 gehört.

Außerdem vermehrt der Prozeß 500 diese Einträge 412 um Vektorinformationen 238A über jeden der zwei oder mehr magischen Vektoren in jedem der schiefwinkligen lokalen Koordinatensysteme 245, die durch jeden der Indizes 414 in der gesamten Datenbank D von molekularen Strukturen dargestellt sind. In diesen Einträgen 412 kann auch zusätzliche Information enthalten sein.

Sobald der Prozeß 500 die Datenstruktur 400 bevölkert hat, enthält die Datenstruktur 400 alle Strukturen (200, 250) und/oder Substrukturen (210, 220, 230) in der gesamten Datenbank D, die gemäß den Tupeleigenschaften klassifiziert sind, die zur Bestimmung des Index 414 verwendet werden, zusammen mit invarianter Information über die magischen Vektoren 238 und 248 (diese können realen oder fiktiven Bindungen entsprechen), die in jenen Strukturen (200, 250) vorhanden sind, und möglicherweise weiteren Informationen.

Fig. 5 ist ein Flußdiagramm, das die Schritte des Bevölkerns der Datenstruktur von Fig. 4 zeigt, damit diese strukturelle Informationen und andere Informationen über ein oder mehrere Referenzmoleküle enthält. Dieser Prozeß wird Referenzspeicherprozeß 500 genannt. Der Prozeß 500 verknüpft einen Index 414, der zu einem Tupel (typischerweise 335, 345, 355) gehört, mit Vektorinformation 420, die den Darstellungen 238A für jeden der zwei oder mehr magischen Vektoren, die mit einer starren Substruktur verknüpft sind, in dem schiefwinkligen lokalen Koordinatensystem 245 des Tupels entspricht, das den Index 414 für jedes Molekül in der Datenbank D aus einer Mehrzahl bekannter Moleküle erzeugt.

Der Prozeß 500 beginnt mit der Auswahl 505 eines Moleküls mit einer Identifikation aus der Datenbank D aus bekannten Molekülen. Diese Identifikation kann eine beliebige bekannte Weise des Bezeichnens eines Moleküls sein, wie oben beschrieben, z. B. ein Schema zur Numerierung der Moleküle.

Schritt 510 bestimmt die Anzahl von starren Substrukturen (210, 220, 230) in dem ausgewählten Molekül 505 und die Anzahl drehbarer Bindungen, die von jeder Substruktur ausgehen.

Dann wird eine starre Substruktur aus dem Satz der starren Substrukturen (210, 220, 230) des ausgewählten Moleküls 505 ausgewählt 515; nachfolgend wird ein Satz von zwei oder mehr magischen Vektoren 238, 248 für die ausgewählte Substruktur bestimmt 520; die Bestimmung der magischen Vektoren und von deren Position und Orientierung in dem globalen Koordinatensystem 235 wird in der früher beschriebenen Weise erreicht. Wie bereits angegeben, kann ein magischer Vektor in Form einer drehbaren Bindung definiert werden, die von der betrachteten Substruktur ausgeht, es ist jedoch nicht notwendig. Es wird ein Referenz-tupel-Auswahlsatz für die ausgewählte 515 starre Substruktur erzeugt 525.

In Schritten 530, 535, 540, 545, 550 und 555 werden ein Tupel, das zugehörige schiefwinklige lokale Koordinatensystem 245 und ein Index (= Referenzkoordinatensystemtupelindex) in der Datenstruktur 400 — wobei der Index eindeutig für das Tupel ist — für jedes Tupel erzeugt, das aus dem Referenz-tupel-Auswahlsatz gebildet werden kann. Bei einer bevorzugten Ausführungsform werden lediglich normierte Tupel verwendet (siehe oben).

In Schritt 530 wird ein Tupel durch Auswählen unter den Elementen des Referenz-tupel-Auswahlsatzes erzeugt. In Schritt 535 wird ein schiefwinkliges lokales Koordinatensystem 245 von dem in 530 erzeugten Tupel erzeugt, wie in den Fig. 2 und 3 oben beschrieben; jeder der zwei oder mehr magischen Vektoren 238, 248, der mit der ausgewählten 515 Substruktur 210, 220, 230 verknüpft ist, wird in dem schiefwinkligen lokalen Koordinatensystem 245, das durch das Tupel definiert ist, dargestellt 540. Oben sind verschiedene Weisen der Darstellung 540 der magischen Vektoren beschrieben.

In Schritt 545 wird der mit dem erzeugten Tupel 530 verknüpfte Index 414 erzeugt (siehe oben hinsichtlich bevorzugter Ausführungsformen zur Erzeugung von Indizes). In Schritt 550 wird die Darstellung 540 jeder der zwei oder mehr magischen Vektoren 238, 248 in dem Datenfeld/der Datenstruktur 400 als ein Eintrag 412 gespeichert. Man beachte, daß der Eintrag 412 mit dem Index 414 verknüpft ist, der dem ausgewählten/erzeugten Tupel 530 entspricht. In Schritt 555 bestimmt der Prozeß 500, ob mehr Tupel von den Elementen des Referenz-tupel-Auswahlsatzes 525 zu erzeugen sind 530. Wenn mehr Tupel zu erzeugen sind, werden die Schritte 530, 535, 540, 545, 550 und 555 wiederholt. Wenn keine Tupel mehr zu erzeugen sind 555, wird das identifizierte Molekül 505 überprüft 560, um zu bestimmen, ob alle seiner starren Substrukturen bearbeitet wurden — mit 'Bearbeitung' ist hier gemeint, daß ein Eintrag 412 in die Datenstruktur 400 gemacht wird. Wenn eine der Substrukturen des identifizierten Moleküls 505 weiterhin unbearbeitet bleibt 560, wird die unbearbeitete Sub-

struktur ausgewählt 515, und die Schritte 520, 525, 530, 535, 540, 545, 550 und 555 werden wiederholt.

Wenn alle starren Substrukturen in dem ausgewählten Molekül 505 bearbeitet wurden, bestimmt 570 der Prozeß 500, ob es irgendwelche unbearbeiteten Moleküle in der Datenbank D gibt. Wenn es welche gibt, beginnt der Prozeß 500 von neuem mit Schritt 505 mit einem neu ausgewählten Molekül. Wenn es keine gibt, endet 575 der Prozeß 500, wobei er die Datenstruktur 400 mit allen möglichen Darstellungen 412 von jedem der zwei oder mehr magischen Vektoren 238, 248 in allen schiefwinkligen lokalen Koordinatensystemen 245 von allen starren Substrukturen 210, 220, 230 jedes Moleküls 505 in der Datenbank D bevölkert hat. Es ist zu erwähnen, daß mehr als eine Darstellung von magischen Vektoren (z. B. 412A bis 412N) in der Datenstruktur 400 angeordnet werden kann, wie sie mit einem gegebenen Index 414 verknüpft ist, der einen Datensatz 425 der Datenstruktur 400 identifiziert.

Fig. 6 ist ein Flußdiagramm, das die Schritte des Vergleichsprozesses 600 zeigt. Der Vergleichsprozess verwendet die Datenstruktur 400, die durch den Referenzspeicherprozeß 500 bevölkert wurde.

Der Prozeß 600 bildet Tupel aus dem Vergleichstupel-Auswahlsatz eines beliebigen gegebenen Testmoleküls und einen Satz von Indizes 414, die diesen Tupeln in der oben beschriebenen Weise entsprechen. Dieser Satz von Indizes ist der 'Testindex'-Satz. Bei gegebener Information in der Datenstruktur 400 und gegebenem Testindex-Satz kann der Prozeß 600 jene Strukturen (200, 250) und/oder Substrukturen (210, 220, 230) aller der Moleküle in der Datenbank bestimmen, die Tupel enthalten, welche identische Eigenschaften A_i mit den Tupeln des Testmoleküls teilen, die zur Erzeugung des Testindex-Satzes verwendet wurden. Des weiteren kann der Prozeß 600 unter Verwendung einer markierenden Datenstruktur 700 und der Information (410, 420) bestimmen, ob das ganze oder ein Teil des Testmoleküls identisch mit einer oder mehr Strukturen (200, 250) und/oder Substrukturen (210, 220, 230) in der Datenbank ist.

Der Prozeß 600 beginnt mit der Auswahl 605 eines Testmoleküls von einer Sammlung aus einem oder mehreren Testmolekülen. Dieses Testmolekül wird gegenüber der Datenbank D geprüft, um jene Moleküle von D, die molekulare Substrukturen (210, 220) enthalten, die mit dem Testmolekül zusammenpassen, zusammen mit dem Satz starrer Transformationen (d. h. starrer Rotationen und Translationen) zu identifizieren, die bewirken, daß jedes dieser Moleküle am besten mit dem ausgewählten 605 Testmolekül überlappt ('beste Übereinstimmung'). Mit 'Zusammenpassen' ist hier gemeint, daß: (a) das identifizierte Molekül (oder die identifizierten Moleküle) in D identisch mit dem Testmolekül ist (sind); oder (b) das identifizierte Molekül (oder die identifizierten Moleküle) in D Substrukturen (210, 220, 230) enthält (enthalten), deren Teile mit dem Testmolekül in seiner Gesamtheit zusammenpassen; oder (c) das Testmolekül einen Teil enthält, der mit dem identifizierten Molekül (oder den identifizierten Molekülen) in D in seiner Gesamtheit zusammenpaßt; oder (d) das Testmolekül einen Teil enthält, der mit Teilen von einer oder mehreren starren Substrukturen in dem identifizierten Molekül (oder den identifizierten Molekülen) von D zusammenpaßt. Man beachte, daß das Testmolekül und das identifizierte Molekül (oder die identifizierten Moleküle) von D nicht in der gleichen Konformation vorliegen müssen. Der Prozeß 600 bestimmt im wesentlichen, ob das Testmolekül mit einem oder mehreren Molekülen in D für eine gegebene Konformation der letzteren zusammenpaßt. Der Prozeß 600 bestimmt außerdem die erforderlichen starren Transformationen, die das (die) identifizierte(n) Molekül(e) in die Konformation bringen, die am besten mit dem Testmolekül übereinstimmt. Für ein verwandtes System und Verfahren, das lediglich die Identität eines (von) Moleküls (Molekülen) in D, jedoch nicht die starren Transformationen bestimmt, die notwendig sind, um es (sie) in beste Übereinstimmung mit dem ausgewählten 605 Testmolekül zu bringen, nehme man bitte Bezug auf die US-Patentanmeldung, eingereicht am 22. Dezember 1995, mit dem Titel 'System and Method for Conformationally Flexible Molecular Identification', von I. Rigoutsos und A. Califano, die am gleichen Tag wie diese Anmeldung eingereicht wurde und die in ihrer Gesamtheit hierin aufgenommen wird.

In dem optionalen Schritt 610 bestimmt der Prozeß 600 unter Verwendung irgendeiner der üblichen Vorgehensweisen, ob eine oder mehrere drehbare Bindungen in dem Testmolekül vorliegen. Indem dies durchgeführt wird, werden alle starren Substrukturen (210, 220, 230) in dem Testmolekül identifiziert.

Wenn lediglich eine starre Substruktur vorhanden ist, wird jene starre Substruktur ausgewählt 620. Wenn es mehr als eine Substruktur gibt, wird eine Substruktur ausgewählt 620, die vorher nicht ausgewählt wurde. Es wird ein Vergleichstupel-Auswahlsatz für die ausgewählte Substruktur des Testmoleküls 605 erzeugt 625.

In Schritten 630, 635, 645 werden ein Tupel, das zugehörige schiefwinklige lokale Koordinatensystem 245 und ein Index (= Testkoordinatensystemtupelindex), der für das Tupel eindeutig ist, für jedes Tupel erzeugt, das von dem Vergleichstupel-Auswahlsatz gebildet werden kann. Bei einer bevorzugten Ausführungsform werden lediglich normierte Tupel verwendet (siehe oben).

In Schritt 630 wird ein Tupel durch Auswahl unter den Elementen des Vergleichstupel-Auswahlsatzes ausgewählt. In Schritt 635 wird ein schiefwinkliges lokales Koordinatensystem 245 von dem in 630 erzeugten Tupel erzeugt, wie in den Fig. 2 und 3 oben beschrieben. In Schritt 645 wird der Testkoordinatensystemtupelindex 645i, der mit dem erzeugten Tupel 630 verknüpft ist, erzeugt (siehe oben hinsichtlich bevorzugter Ausführungsformen zur Erzeugung von Indizes).

Man beachte, daß die Schritte 610, 615, 620, 625, 630, 635 und 645 für das Testmolekül in einer identischen Weise durchgeführt werden wie die jeweiligen Schritte 510, 515, 520, 525, 530, 535 und 545 für alle Referenzmoleküle in der Datenbank D durch den Prozeß 500 durchgeführt werden. Daher ist der Testkoordinatensystemtupelindex 645i für das zugehörige Tupel eindeutig und invariant unter Translation 295 und Rotationen 290 der molekularen Struktur (200, 250) und jeglicher Rotationen 215 einer beliebigen Substruktur (210, 220, 230) um drehbare Bindungen 218a, 218b herum, die in dem ausgewählten Molekül 605 vorhanden sind.

In Schritt 650 gewinnt der Prozeß 600 Darstellungen und andere Informationen aus der Datenstruktur (dem Datenfeld) 400 unter Verwendung des Testkoordinatensystemtupelindex. In dem Fall, in dem das Testmolekül identisch (in allen Gesichtspunkten, die durch den gebildeten Index abgedeckt sind; z. B. physikalische, chemische, geometrische etc.) mit einem oder mehreren der Moleküle in der Datenbank D ist, gibt es wenigstens einen

Eintrag 412 von Vektorinformation 420 in dem Datensatz 425, auf den durch jeden erzeugten Testkoordinatensystemtupelindex 645i zugegriffen wird, in der Datenstruktur 400, der die gleiche Vektorinformation aufweist, die jeden von zwei oder mehr magischen Vektoren in dem Testmolekül beschreibt. Der Testkoordinatensystemtupelindex 645i greift auf den Datensatz 425 zu, da der Testkoordinatensystemtupelindex 645i identisch mit dem Referenzkoordinatensystemtupelindex 414 ist, da sie beide von den gleichen molekularen Substrukturen (210, 220, 230) unter Verwendung der gleichen Schritte (510, 515, 520, 525, 530, 535, 545 beziehungsweise 610, 615, 620, 625, 630, 635, 645) erzeugt wurden.

Es ist jedoch zu erwähnen, daß es weitere Moleküle (oder starre Substrukturen und/oder Teile von starren Substrukturen) in der Datenbank D geben kann, die Tupel enthalten, die Referenzkoordinatensystemtupelindizes 414 erzeugen, welche die gleichen wie die Testkoordinatensystemtupelindizes 645i sind. Dies ist so, da die entsprechenden Tupel identisch in bezug auf die gewählten Eigenschaften A_i sind, aus denen sowohl der Referenzkoordinatensystemtupelindex 414 als auch der Testkoordinatensystemtupelindex 645i besteht. Zum Beispiel erzeugt in dem Fall, in dem die Eigenschaften geometrisch (11/12/13, wie oben) und vom Atomtyp eines Platzes (Atomtyp, wie oben) sind, das Tupel A-B-E in Fig. 2A den gleichen Index, ungeachtet des tatsächlichen chemischen Typs der, Atome B und E, solange die Werte der Eigenschaften, die den Index bilden, identisch bleiben. Daher weist die Struktur 400 Informationen auf, die beim Identifizieren von einem oder mehr Molekülen (ebenso wie der notwendigen starren Transformationen) von der Datenbank D, die mit einem gegebenen Testmolekül zusammenpassen (siehe oben hinsichtlich einer Definition von 'Zusammenpassen') durch Bestimmen der Häufigkeit des Auftretens impliziter oder expliziter Information nützlich sind, die durch die Vektorinformation 420 in einer oder mehr der Einträge 412A bis 412N gegeben ist, wie unten beschrieben.

Nach der Gewinnung der Vektorinformation für die drehbaren Bindungen in Schritt 650 wird die Vektorinformation 420 für jeden Eintrag 412A bis 412N des Datensatzes 425, auf den durch den Testkoordinatensystemtupelindex 645i zugegriffen wird, dazu verwendet, die Position und Orientierung in dem globalen Koordinatensystem 235 von jedem der zwei oder mehr magischen Vektoren zu gewinnen, die in jedem Eintrag 412A bis 412N in dem Datensatz 425 enthalten sind. Diese gewonnenen Fälle der magischen Vektoren können in dieser Erörterung auch als Testvektoren bezeichnet werden. Die Gewinnung wird durch Verwenden der Darstellungen von jedem der zwei oder mehr magischen Vektoren, die in den Einträgen 412A bis 412N enthalten sind, und übliche Vektoranalyseverfahren erreicht; für jeden Eintrag in jedem Datensatz mit einem Referenzkoordinatensystemtupelindex, der mit dem Testkoordinatensystemtupelindex zusammenpaßt, erzeugen wir einen Abstimm-datensatz in einer Abstimmdatenstruktur 655, wobei der Abstimm-datensatz Anordnungsinformationen in dem globalen Koordinatensystem 235 für jeden der magischen Vektoren enthält, deren Darstellungen in den Einträgen 412A bis 412N enthalten sind. Bei alternativen bevorzugten Ausführungsformen können die molekulare Identität 421A bis 421N und/oder Substruktur (210, 220, 230)-Identität 422A bis 422N zusätzlich zu der gewonnenen Anordnungsinformation verwendet werden, wenn die Abstimmtablelle bevölkert wird.

In Schritt 660 wird jeder der in Schritt 650 erzeugten Abstimm-datensätze in die Abstimmtablelle eingegeben (siehe 700 unten). Klarerweise erzeugt Schritt 650 viele identische Abstimm-datensätze, d. h. Abstimm-datensätze, welche die gleiche Anordnungsinformation, molekulare Identitätsinformation und Substrukturidentitätsinformation enthalten. Dies ist das Ergebnis von mehr als einem Koordinatensystemtupel, die eine bestimmte Anordnung in dem globalen Koordinatensystem 235 für eine gegebene starre Substruktur eines Moleküls erhärten. Das Ausmaß an Zusammenpassen zwischen einem Teil eines Testmoleküls und einem oder mehr Teilen von einer oder mehr Substrukturen von einem oder mehr Molekülen in der Datenbank D steht in direkter Beziehung zu der Multiplizität derartiger identischer Abstimm-datensätze oder äquivalent zu der Häufigkeit des Auftretens jedes der unterschiedlichen Abstimm-datensätze in der Abstimmtablelle 700.

Sobald alle Abstimm-datensätze, die unter Verwendung der Punkte in den Einträgen 412A bis 412N der Vektorinformation 420 für den Datensatz 425, auf den zugegriffen wurde, erzeugt wurden, in die Abstimmtablelle eingegeben sind, bestimmt 665 dann der Prozeß 600, ob mehr Tupel von den Elementen des Referenz-tupel-Auswahlsatzes 625 zu erzeugen sind 630. Wenn mehr Tupel zu erzeugen sind 665, werden die Schritte 630, 635, 645, 650 und 655 wiederholt. Wenn keine Tupel mehr zu erzeugen sind 665, wird das Testmolekül 605 geprüft 670, um zu bestimmen, ob alle Substrukturen (210, 220, 230) bearbeitet wurden.

Wenn es noch mehr unbearbeitete Substrukturen gibt, wird eine derartige Substruktur ausgewählt 620, und die Schritte 625, 630, 635, 645, 650 und 655 werden wiederholt.

Nach Beendigung der Bearbeitung des ausgewählten Testmoleküls 605 wurde die Abstimmtablelle 700, die in Fig. 7 gezeigt ist, durch Abstimm-datensätze 725 bevölkert, die durch die Einträge der Datenstruktur 400 erzeugt wurden.

Jeder Datensatz 725 der Abstimmtablelle besitzt eine Adresse 710 und enthält die Information über die Referenzmolekülidentität, die Information über die Identität der starren Referenzkoordinatensystems-substruktur und Anordnungsinformationen für jeden der zwei oder mehr magischen Vektoren, deren Darstellungen in Einträgen 412A bis 412N des Datensatzes 425 enthalten sind, auf den durch den Testkoordinatensystemtupelindex 465i zugegriffen wird.

Bei einer bevorzugten Ausführungsform werden die Molekülidentität 736 und/oder die Identität 738 der starren Substruktur 210 dazu verwendet, um die Adresse 710 jedes Abstimm-datensatzes 725 zu berechnen. Die Adresse 710 wird durch das oben beschriebene 'Schritt'-Berechnungsverfahren bestimmt. Bei einer alternativen bevorzugten Ausführungsform können die Anordnungsinformationen für jeden der zwei oder mehr magischen Vektoren, deren Darstellungen in den Einträgen 412A bis 412N des Datensatzes 425 enthalten sind, dazu verwendet werden, die Adresse 710 des Datensatzes 725 abzuleiten.

Nunmehr zu Fig. 6 zurückkehrend, wird die bevölkerte Abstimmtablelle 700 dazu verwendet, um zu bestimmen: (i) die Identität von einem oder mehr Molekülen in der Datenbank D, (ii) die Identität von einer oder mehr starren Substrukturen in jedem Molekül und (iii) die Position und Orientierung der mit jeder starren Substruktur

verknüpften magischen Vektoren, so daß (a) eine starre Substruktur in jedem derartigen Molekül der beste Kandidat für ein Zusammenpassen mit einer Substruktur in dem Testmolekül ist, und (b) wenn eine derartige starre Substruktur in dem globalen Koordinatensystem 235 angeordnet ist, so daß die Position und Orientierung der zugehörigen magischen Vektoren mit denjenigen zusammenpaßt, die in (iii) bestimmt wurden, jedes identifizierte Molekül in bester Übereinstimmung mit dem Testmolekül ist. Man beachte, daß es mehr als ein Molekül in der Datenbank D geben kann, die beste Kandidaten für ein Zusammenpassen mit einer Substruktur in dem Testmolekül sind, und dies ist eine Folge davon, daß eine gegebene Testmolekülsubstruktur von mehr als einem Molekül in der Datenbank D geteilt werden kann. Eine Bestimmung dieser Antworten (i), (ii) und (iii) kann durch Unterauswahl jener Datensätze aus der Abstimmtable 700 mit einem Zählwert (einer Häufigkeit) bewerkstelligt werden, der (die) einen vorgegebenen Schwellwert 675 übersteigt. Diese ausgewählten Datensätze 725 repräsentieren die rekonstruierten P-unkte mit den Eigenschaften (a) und, (b) oben.

Hin und wieder kann es wünschenswert sein, die hypothetischen Anordnungen in jenen aus der Abstimmtable 700 erhaltenen Antworten zu verwenden, die sich auf das gleiche Molekül aus der Datenbank D beziehen, um jene Konformation des Moleküls zu bilden, die in Übereinstimmung mit so vielen dieser Antworten wie möglich ist. Wenn das fragliche Molekül in diese Konformation gebracht wird, befindet es sich in seiner bestmöglichen Übereinstimmung mit dem Testmolekül als ganzem. Die Qualität der Übereinstimmung zwischen den zwei Molekülen variiert als Funktion des tatsächlichen Grades an Ähnlichkeit zwischen ihnen, wenn alle Konformationen des identifizierten Moleküls berücksichtigt werden. Diese Kombination von Antworten kann mit minimalem Berechnungsaufwand erreicht werden: jede Antwort enthält bereits Informationen über die Anordnung der magischen Vektoren, die mit der bestimmten starren Substruktur verknüpft sind; da die magischen Vektoren konstruktionsbedingt starr an der jeweiligen starren Substruktur angebunden sind, plaziert ihre Anordnung in dem globalen Koordinatensystem 235 unmittelbar die starre Substruktur und alle drehbaren Bindungen, die von der Substruktur ausgehen, in dem gleichen Koordinatensystem. Zwei Antworten, die dem gleichen Molekül der Datenbank D, jedoch verschiedenen starren Substrukturen entsprechen, die über eine drehbare Bindung verbunden sind, können zu einer 'Zwei-Substruktur'-Teilantwort miteinander kombiniert werden, wenn die Anordnung der jeweiligen starren Substrukturen mit der Plazierung der gemeinsamen drehbaren Bindung, die sie verbindet, übereinstimmt; eine Zwei-Substruktur-Teilantwort kann des weiteren zu einer 'Drei-Substruktur'-Teilantwort unter Verwendung der hypothetischen Plazierung einer dritten starren Substruktur indem betrachteten Molekül erweitert werden, wenn die dritte Substruktur mit jeder der ersten zwei Substrukturen über eine drehbare Bindung verbunden ist und ihre Anordnung in dem globalen Koordinatensystem die gemeinsame drehbare Bindung in der Position und Orientierung plaziert, die von dem Zwei-Substruktur-Komplex erfordert wird. Klarerweise kann dieser Prozeß mit dem neu gebildeten Drei-Substruktur-Komplex fortgesetzt werden und so weiter und so fort. Wenn die Antworten, die den vorgegebenen Schwellwert überschreiten, einen Subsatz enthalten, der sich auf das gleiche Molekül bezieht, kann es möglich sein, alle starren Substrukturen des Moleküls zu berücksichtigen und sie in einer global konsistenten Konformation zu plazieren. Typischerweise erzeugen die konsistenten Antworten Komplexe starrer Substrukturen, die eine oder mehr Substrukturen von jedem der identifizierten Moleküle aus der Datenbank D beinhalten.

Die von der Tabelle 700 erhaltenen Antworten benötigen einen minimalen Berechnungsaufwand, um jene Konformationen zu erzeugen, welche die identifizierten Moleküle aus der Datenbank D in bester Übereinstimmung mit dem Testmolekül anordnen.

In Anbetracht dieser Offenbarung kann der Fachmann äquivalente alternative Ausführungsformen für die Identifikation von Molekülen entwickeln, die ebenfalls innerhalb der Erwägungen der Erfinder liegen.

Patentansprüche

1. Verfahren zum Speichern einer Darstellung eines oder mehrerer Referenzmoleküle in einem Speicher in einem Rechnersystem, wobei das Verfahren auf einem Rechnersystem durchgeführt wird und die Schritte umfaßt:

- a. Erkennen entweder einer oder mehrerer starrer Substrukturen des Referenzmoleküls, wobei jede der starren Substrukturen einen oder mehr Atomplätze aufweist, jeder der Atomplätze mit null oder mehr Atomplätzen in der starren Substruktur über eine nicht drehbare Bindung verbunden ist und jede starre Substruktur eine globale Position und eine globale Orientierung in einem globalen Koordinatensystem besitzt;
- b. Definieren von zwei oder mehr Vektoren mit einer Größe und Richtung mit einer festen Position und Orientierung bezüglich einer ausgewählten starren Substruktur, wobei die ausgewählte starre Substruktur eine der starren Substrukturen ist;
- c. Auswählen eines Satzes von drei oder mehr Plätzen, wobei sich der ausgewählte Satz von Plätzen in der ausgewählten starren Substruktur befindet, der Satz von Plätzen ein Koordinatensystemtupel bildet, wenigstens einer der Plätze nicht kollinear mit den restlichen Plätzen ist, die Plätze sich in einer festen Position bezüglich der ausgewählten starren Substruktur befinden und das Koordinatensystemtupel ein dreidimensionales schiefwinkliges lokales Koordinatensystem definiert;
- d. Auswählen eines oder mehr der Koordinatensystemtupel und Erzeugen eines Koordinatensystemtupelfeldes mit Informationen, die mit jedem der ausgewählten Koordinatensystemtupel verknüpft sind; und
- e. Speichern eines Datensatzes in einer Datenstruktur, wobei die Datenstruktur eine Mehrzahl von Datensätzen aufweist, jeder Datensatz das Koordinatensystemtupelfeld und ein Vektorfeld enthält und das Vektorfeld Vektorinformationen, die sich auf jeden der Vektoren beziehen, sowie Informationen über die Identitäten des Moleküls und der ausgewählten starren Substruktur enthält.

2. Verfahren nach Anspruch 1, wobei die Information in dem Koordinatensystemtupelfeld ein Index ist.
3. Verfahren nach Anspruch 2, wobei der Index von geometrischen Informationen abgeleitet wird, die sich auf den Satz von Plätzen beziehen.
4. Verfahren nach Anspruch 3, wobei der Index von einem oder mehreren der Abstände zwischen jeweils zwei Plätzen in dem Satz von Plätzen abgeleitet wird.
5. Verfahren nach Anspruch 3, wobei der Index von einem oder mehr der Winkel in einem oder mehr der Dreiecke abgeleitet wird, die durch jeweils drei Plätze in dem Satz von Plätzen gebildet werden.
6. Verfahren nach Anspruch 3, wobei der Index von einer Kombination von null oder mehr Winkeln in einem oder mehr der Dreiecke, die durch jeweils drei Plätze in dem Satz von Plätzen gebildet werden, und null oder mehr Abständen zwischen jeweils zwei Plätzen des Satzes von Plätzen abgeleitet wird.
7. Verfahren nach Anspruch 2, wobei der Index von physikalischer Information abgeleitet wird, die für einen oder mehr der Plätze des Koordinatensystemtupels charakteristisch ist.
8. Verfahren nach Anspruch 2, wobei der Index von chemischer Information abgeleitet wird, die für einen oder mehr der Plätze des Koordinatensystemtupels charakteristisch ist.
9. Verfahren nach Anspruch 2, wobei der Index von geometrischer Information, die sich auf null oder mehr der Plätze des Koordinatensystemtupels bezieht, physikalischer Information, die für null oder mehr der Plätze des Koordinatensystemtupels charakteristisch ist, und chemischer Information abgeleitet wird, die für null oder mehr der Plätze des Koordinatensystemtupels charakteristisch ist.
10. Verfahren nach Anspruch 1, wobei die Vektorinformation jeden der Vektoren in dem schiefwinkligen lokalen Koordinatensystem eindeutig identifiziert und die Vektorinformation unter jeglicher Rotation und Translation des Satzes von Plätzen, der das schiefwinklige lokale Koordinatensystem definiert, invariant bleibt.
11. Verfahren nach Anspruch 10, wobei die Vektorinformation Information über eine Identität, eine Position, eine Vektorgröße und eine Vektororientierung von jedem der Vektoren ist, die in dem lokalen schiefwinkligen Koordinatensystem dargestellt werden.
12. Verfahren nach Anspruch 11, wobei die Vektorinformation die Projektion von jedem der Vektoren auf eine oder mehr Achsen des lokalen schiefwinkligen Koordinatensystems enthält.
13. Verfahren nach Anspruch 11, wobei einer oder mehr der Vektoren durch zwei oder mehr Plätze dargestellt wird, wobei die Plätze Atomplätze des Moleküls sind, das einen ersten und einen zweiten Atomplatz beinhaltet, und der erste und der zweite Atomplatz die Position, Größe und Orientierung des jeweiligen Vektors definieren.
14. Verfahren nach Anspruch 11, wobei einer oder mehr der Vektoren durch eine Punktposition eines festen Punktes entlang der Länge des Vektors, die Vektorgröße und die Vektororientierung repräsentiert wird.
15. Verfahren nach Anspruch 11, wobei einer oder mehr der Vektoren durch die Position repräsentiert wird, wobei die Position durch zwei Vektorplätze bestimmt wird, die Vektorplätze Plätze in dem Satz von Plätzen sind und die Vektorinformation des weiteren die Größe und die Orientierung des Vektors beinhaltet.
16. Verfahren nach Anspruch 15, wobei einer oder mehr der Vektorplätze ein Atomplatz ist.
17. Verfahren nach Anspruch 16, wobei einer oder mehr der Vektorplätze ein Nicht-Atomplatz ist.
18. Verfahren nach Anspruch 11, wobei die Position, Größe und Orientierung von einem oder mehr der Vektoren durch eine Matrix dargestellt werden.
19. Verfahren nach Anspruch 2, wobei die Vektorinformation des weiteren andere Informationen beinhaltet.
20. Verfahren nach Anspruch 19, wobei die anderen Informationen irgendeine oder mehr der folgenden beinhalten: molekulare Identität, Identität der Substruktur, Information über Atomplätze und Information über Nicht-Atomplätze sowie Information über die Kardinalität und Identität von einem oder mehr der Vektoren.
21. Verfahren nach Anspruch 20, wobei die anderen Informationen des weiteren irgendeine oder mehr der folgenden beinhalten:
physikalische Eigenschaften und chemische Eigenschaften.
22. Verfahren nach Anspruch 1, wobei einer oder mehr der Plätze in dem Satz von Plätzen ein Atomplatz ist, der starr mit der ausgewählten starren Substruktur verbunden ist.
23. Verfahren nach Anspruch 1, wobei einer oder mehr der Plätze ein Nicht-Atomplatz ist.
24. Verfahren nach Anspruch 1, wobei einer der Plätze ein erster Platz auf der ausgewählten starren Substruktur ist und ein weiterer Platz ein zweiter Platz auf einer zweiten starren Substruktur ist und wobei die ausgewählte und die zweite starre Substruktur über eine drehbare Bindung verbunden sind.
25. Verfahren nach Anspruch 24, wobei der erste und der zweite Platz irgendeiner der folgenden sein kann: ein Atomplatz und ein Nicht-Atomplatz.
26. Verfahren nach Anspruch 25, wobei der erste Platz ein erster Atomplatz ist, der zweite Platz ein zweiter Atomplatz ist und einer oder mehr der Vektoren die Position, Größe und Orientierung einer drehbaren Bindung hat, die den ersten und den zweiten Atomplatz verbindet.
27. Verfahren nach Anspruch 1, das den weiteren Schritt umfaßt:
f. Wiederholen der Schritte d und e für eines oder mehr der nicht ausgewählten Koordinatensystemtupel.
28. Verfahren nach Anspruch 27, wobei die Schritte d und e für alle nicht ausgewählten Koordinatensystemtupel wiederholt werden.
29. Verfahren nach Anspruch 27, das den weiteren Schritt umfaßt: g. Wiederholen der Schritte c bis f für einen oder mehr der nicht ausgewählten Sätze von Plätzen.
30. Verfahren nach Anspruch 29, wobei die Schritte c bis f für alle nicht ausgewählten Sätze von Plätzen wiederholt werden.
31. Verfahren nach Anspruch 29, das den weiteren Schritt umfaßt: h. Wiederholen der Schritte b bis g für

eine oder mehr der restlichen starren Substrukturen.

32. Verfahren nach Anspruch 31, wobei die Schritte b bis g für alle restlichen starren Substrukturen wiederholt werden.

33. Verfahren nach Anspruch 31, das den weiteren Schritt umfaßt: i. Wiederholen der Schritte a bis h für ein oder mehr nicht ausgewählte Moleküle.

34. Verfahren nach Anspruch 33, wobei die Schritte a bis i für alle nicht ausgewählten Moleküle wiederholt werden.

35. Verfahren zum Speichern einer Darstellung von einem oder mehreren Referenzmolekülen im Speicher eines Rechnersystems, wobei das Verfahren auf einem Rechnersystem durchgeführt wird und die Schritte umfaßt:

a. Bestimmen einer oder mehrerer starrer Substrukturen des Referenzmoleküls, wobei jede der starren Substrukturen einen oder mehr Atomplätze besitzt, jeder der Atomplätze mit null oder mehr Atomplätzen durch eine nicht drehbare Bindung verbunden ist und jede starre Substruktur eine globale Position und eine globale Orientierung in einem globalen Koordinatensystem besitzt;

b. Definieren von zwei oder mehreren Referenzvektoren, jeder mit einer Vektorgröße und einer Vektorrichtung, wobei die Vektoren hinsichtlich Position und Orientierung in bezug auf die ausgewählte starre Substruktur fest sind;

c. Auswählen eines Satzes von drei oder mehr Plätzen, wobei die Plätze sich in einer festen Position bezüglich der ausgewählten starren Substruktur befinden und jeder beliebige Satz von Plätzen ein Koordinatensystemtupel ist, das ein schiefwinkliges lokales Koordinatensystem definiert, wobei das schiefwinklige lokale Koordinatensystem zwei oder mehr Seiten mit einem Winkel zwischen einem oder mehr Paaren der Seiten besitzt;

d. Auswählen von einem oder mehr Koordinatensystemtupeln und Erzeugen von einem oder mehr Indizes von Informationen über jedes der ausgewählten Koordinatensystemtupel; und

e. Speichern eines Datensatzes in einer Datenstruktur, die in dem Speicher gespeichert ist, wobei die Datenstruktur eine Mehrzahl von Datensätzen aufweist, wobei jeder der Datensätze Vektorinformationen über jeden der Referenzvektoren enthält und der Datensatz mit einem der Indizes verknüpft ist und auf ihn durch Verwenden des Index zugegriffen werden kann.

36. Verfahren nach Anspruch 35, wobei der Index aus Informationen von zwei der Seiten erzeugt wird, die eine erste Seite und eine zweite Seite darstellen, wobei ein erster Winkel der Winkel zwischen der ersten und der zweiten Seite ist.

37. Verfahren nach Anspruch 36, wobei entweder die erste oder die zweite Seite eine Größe oberhalb eines Längenschwellwerts besitzt.

38. Verfahren nach Anspruch 36, wobei die erste Seite die längste Seite in dem Dreieck ist, das durch die drei Mitglieder des Koordinatensystemtupels erzeugt wird, und die zweite Seite die zweitlängste Seite des Dreiecks ist, das durch die drei Mitglieder des Koordinatensystemtupels erzeugt wird.

39. Verfahren nach Anspruch 38, wobei der erste Winkel eine Größe oberhalb eines Winkelschwellwerts besitzt.

40. Verfahren nach Anspruch 39, wobei der Winkelschwellwert 10 Grad beträgt.

41. Verfahren nach Anspruch 35, wobei der erste Winkel der größte Winkel in dem Dreieck ist, das durch die drei Mitglieder des Koordinatensystemtupels gebildet wird.

42. Verfahren nach Anspruch 35, wobei der Index aus Informationen erzeugt wird, die des weiteren die chemischen Charakteristika von einem oder mehr der Atomplätze beinhalten, die an dem Koordinatensystemtupel teilhaben.

43. Verfahren zum Berichten von Identitäten und eines Satzes von notwendigen starren Transformationen für ein oder mehr Referenzmoleküle, die einem Testmolekül strukturell ähnlich sind, wobei das Verfahren auf einem Rechnersystem durchgeführt wird und die Schritte umfaßt:

a. Bestimmen von einer oder mehreren starren Testsubstrukturen des Testmoleküls, wobei jede der starren Testsubstrukturen einen oder mehr Atomplätze aufweist, jeder der Atomplätze mit null oder mehreren Atomplätzen in der starren Testsubstruktur durch eine nicht drehbare Bindung verbunden ist und jede starre Testsubstruktur eine bestimmte Position und eine bestimmte Orientierung in einem dreidimensionalen globalen Referenzkoordinatensystem besitzt;

b. Auswählen eines Satzes von drei oder mehr Testplätzen, wobei der Satz von Testplätzen ein Testkoordinatensystemtupel ist, wobei wenigstens einer der Testplätze nicht kollinear mit den restlichen Testplätzen ist, die Testplätze sich in einer festen Position in bezug auf die starre Testsubstruktur befinden und jedes der Testkoordinatensystemtupel ein dreidimensionales schiefwinkliges lokales Testkoordinatensystem definiert;

c. Auswählen von einem oder mehr der Testkoordinatensystemtupel und Erzeugen eines Testkoordinatensystemtupelindex aus Informationen, die mit dem ausgewählten Testkoordinatensystemtupel verknüpft sind;

d. Verwenden des Testkoordinatensystemtupelindex, der auf einen oder mehr Datensätze in einer in dem Speicher gespeicherten Datenstruktur zugreift, wobei die Datenstruktur eine Mehrzahl von Datensätzen aufweist, jeder der Datensätze ein Referenzkoordinatensystemtupelfeld und ein Referenzvektorinformationfeld enthält, das Referenzkoordinatensystemtupelfeld einen Referenzkoordinatensystemtupelindex besitzt, der von einem Referenzkoordinatensystemtupel erzeugt wird, das durch drei oder mehr Referenzplätze auf einer starren Referenzsubstruktur eines der Referenzmoleküle definiert wird, wobei das Referenzvektorfeld einen oder mehr Einträge besitzt, jeder Eintrag Referenzvektorinformationen über die zwei oder mehr Referenzvektoren enthält, jeder der Referenzvektoren

ren eine Größe und Richtung und eine feste Position und eine feste Orientierung bezüglich einer oder mehr der starren Referenzsubstrukturen hat, jeder Eintrag des weiteren Referenzkoordinatensystemtupelinformationen über das Referenzkoordinatensystemtupel, Identitätsinformationen über das Referenzmolekül und Informationen über die starre Referenzkoordinatensystemsubstruktur besitzt;

e. Berechnen eines Testvektors für jeden der zwei oder mehr Referenzvektoren in dem schiefwinkligen lokalen Testkoordinatensystem für jeden Eintrag in jedem Datensatz mit einem Referenzkoordinatensystemtupelindex, der mit dem Testkoordinatensystemtupelindex zusammenpaßt, um jeden der Testvektoren in dem globalen Koordinatensystem anzuordnen; und

f. Erzeugen eines Abstimm Datensatzes in einer Abstimm Datenstruktur für jeden Eintrag in jedem Datensatz mit einem Referenzkoordinatensystemtupelindex, der mit dem Testkoordinatensystemtupelindex zusammenpaßt, wobei der Abstimm Datensatz die Identitätsinformation für das Referenzmolekül, die Identitätsinformation für die starre Referenzkoordinatensystemsubstruktur und eine Anordnungsinformation für jeden der Testvektoren in dem globalen Koordinatensystem enthält.

44. Verfahren nach Anspruch 43, wobei der Referenzkoordinatensystemtupelindex aus dem Referenzkoordinatensystemtupel erzeugt wird und der Testkoordinatensystemtupelindex aus dem ausgewählten Testkoordinatensystemtupel durch das gleiche Verfahren erzeugt wird.

45. Verfahren nach Anspruch 43, das den weiteren Schritt umfaßt:

g. Wiederholen der Schritte c bis f für eines oder mehr der nicht ausgewählten Testkoordinatensystemtupel.

46. Verfahren nach Anspruch 45, wobei die Schritte c bis f für alle nicht ausgewählten Testkoordinatensystemtupel wiederholt werden.

47. Verfahren nach Anspruch 45, das den weiteren Schritt umfaßt:

h. Wiederholen der Schritte b bis g für einen oder mehr der nicht ausgewählten Sätze von Testplätzen.

48. Verfahren nach Anspruch 45, wobei die Schritte b bis g für alle nicht ausgewählten Sätze von Plätzen wiederholt werden.

49. Verfahren nach Anspruch 48, das den weiteren Schritt umfaßt:

i. Wiederholen der Schritte a bis g für eine oder mehr der nicht ausgewählten starren Testsstrukturen.

50. Verfahren nach Anspruch 48, wobei die Schritte a bis g für alle der nicht ausgewählten Testsstrukturen wiederholt werden.

51. Verfahren nach Anspruch 49, wobei eine Multiplizität des Auftretens für einen oder mehr identische Abstimm Datensätze für jeden von einem oder mehr Sätzen identischer Abstimm Datensätze bestimmt wird.

52. Verfahren nach Anspruch 51, wobei die Sätze von Abstimm Datensätzen, die einen Multiplizitätswert unterhalb eines Schwellwerts besitzen, eliminiert werden.

53. Verfahren nach Anspruch 51, wobei der Abstimm tabellendatensatz, der die höchste Multiplizität des Auftretens aufweist, die Identität des Referenzmoleküls, das mit dem Testmolekül zusammenpaßt, und die Identität der starren Referenzsubstruktur enthält, die mit einer oder mehr der starren Testmolekülsstrukturen zusammenpaßt, und die Anordnungsinformation die notwendigen starren Transformationen bestimmt, welche die Referenz- und Testmoleküle in Übereinstimmung bringen.

54. Verfahren nach Anspruch 3, wobei der Index von Flächen abgeleitet wird, die durch Verwenden von wenigstens drei Plätzen in dem Subsatz von Plätzen gebildet werden.

55. Verfahren nach Anspruch 3, wobei der Index von Verhältnissen von Flächen abgeleitet wird, die durch Verwenden von wenigstens drei Plätzen in dem Subsatz von Plätzen gebildet werden.

56. Rechnersystem zum Speichern einer Darstellung von einem oder mehr Referenzmolekülen in einem Speicher in dem Rechnersystem und zum Vergleichen von einem oder mehr der Referenzmoleküle mit einem Testmolekül, das beinhaltet:

a. eine Datenbank, die in dem Speicher gespeichert ist, wobei die Datenbank eine Darstellung von einer oder mehr starren Substrukturen von jedem der Referenzmoleküle besitzt, jede der starren Substrukturen einen oder mehr Atomplätze aufweist, jeder der Atomplätze mit null oder mehr Atomplätzen in der starren Substruktur durch eine nicht drehbare Bindung verbunden ist, jede starre Substruktur eine globale Position und eine globale Orientierung in einem globalen Koordinatensystem besitzt;

b. einen Satz von drei oder mehr Plätzen, wobei sich der Satz von Plätzen in einer ausgewählten starren Substruktur befindet, der Satz von Plätzen ein Koordinatensystemtupel bildet, wenigstens einer der Plätze nicht kollinear mit den restlichen Plätzen ist, die Plätze sich in einer festen Position bezüglich der ausgewählten starren Substruktur befinden und das Koordinatensystemtupel ein dreidimensionales schiefwinkliges lokales Koordinatensystem definiert; und

c. eine Datenstruktur mit einer Mehrzahl von Datensätzen, wobei jeder Datensatz ein Koordinatensystemtupelfeld und ein Vektorfeld enthält und das Vektorfeld Vektor-Informationen, die sich auf jeden der zwei oder mehr Vektoren beziehen, sowie Informationen über die Identitäten von einem oder mehr der Moleküle und einer oder mehr der starren Substrukturen enthält, wobei jeder der Vektoren eine Größe und eine Richtung sowie eine feste Position und Orientierung bezüglich der ausgewählten starren Substruktur besitzt und die ausgewählte starre Substruktur eine der starren Substrukturen ist.

57. Rechnersystem nach Anspruch 56, das des weiteren eine Abstimm Datenstruktur aufweist, wobei die Abstimm Datenstruktur eine Mehrzahl von Abstimm Datensätzen besitzt, wobei jeder der Abstimm Datensätze Informationen enthält, welche die Identität eines Referenzmoleküls, die Identität einer starren Referenzkoordinatensystemsubstruktur und eine Anordnungsinformation für jeden der zwei oder mehr Testvektoren in dem globalen Koordinatensystem mit einem Testkoordinatensystemtupelindex umfaßt, welcher der gleiche ist wie ein Referenzkoordinatensystemtupelindex, der mit dem Koordinatensystemtupelfeld und

dem Vektorfeld verknüpft ist, und dem Testkoordinatensystemtupelindex als das Koordinatensystemtupelfeld, wobei der Testkoordinatensystemtupelindex für jedes von dem einen oder mehr ausgewählten Testkoordinatensystemtupel des Testmoleküls und aus Informationen erzeugt wird, die mit dem ausgewählten Testkoordinatensystemtupel verknüpft sind, wobei jedes der ausgewählten Testkoordinatensystemtupel aus einem Satz von drei oder mehr Testplätzen des Testmoleküls gebildet wird, wenigstens einer der Testplätze nicht kollinear mit den restlichen Testplätzen ist, die Testplätze sich in einer festen Position bezüglich der starren Testsstruktur auf dem Testmolekül befinden und jedes der Testkoordinatensystemtupel ein dreidimensionales schiefwinkliges lokales Testkoordinatensystem definiert.

58. System nach Anspruch 57, wobei eine Multiplizität des Auftretens für einen oder mehr identische Abstimmungsätze für jeden der einen oder mehr Sätze von identischen Abstimmungsätzen bestimmt wird.

59. System nach Anspruch 58, wobei die Sätze von Abstimmungsätzen, die einen Multiplizitätswert unterhalb eines Schwellwerts aufweisen, eliminiert werden.

60. Rechnersystem zum Speichern einer Darstellung von einem oder mehreren Referenzmolekülen in einem Speicher in dem Rechnersystem und zum Vergleichen von einem oder mehr der Referenzmoleküle mit einem Testmolekül, das beinhaltet:

a. Datenbankmittel, die in dem Speicher gespeichert sind, wobei die Datenbankmittel eines oder mehr starre Substrukturmittel von jedem der Referenzmoleküle darstellen, jedes, der starren Substrukturmittel ein oder mehr Atomplatzmittel besitzt, jedes der Atomplatzmittel mit null oder mehr Atomplatzmitteln in den starren Substrukturmitteln durch eine nicht drehbare Bindung verbunden ist und jedes starre Substrukturmittel eine globale Position und eine globale Orientierung in einem globalen Koordinatensystem besitzt;

b. einen Satz von drei oder mehr Platzmitteln, wobei der Satz von Plätzen sich in einem ausgewählten starren Substrukturmittel befindet, der Satz von Platzmitteln ein Koordinatensystemtupelmittel bildet, wenigstens eines der Platzmittel nicht kollinear mit den restlichen Platzmitteln ist, die Platzmittel sich in einer festen Position bezüglich der ausgewählten starren Substrukturmittel befinden und die Koordinatensystemtupelmittel ein dreidimensionales schiefwinkliges lokales Koordinatensystemmittel definieren; und

c. ein Datenstrukturmittel zum Speichern einer Mehrzahl von Datensatzmitteln, wobei jedes Datensatzmittel ein Koordinatensystemtupelfeld und ein Vektorfeld enthält, das Vektorfeld Vektorinformationen, die sich auf jeden der zwei oder mehr Vektoren beziehen, sowie Informationen über die Identitäten von einem oder mehr der Moleküle und einem oder mehr der starren Substrukturmittel enthält, wobei jeder der Vektoren eine Größe und eine Richtung mit einer festen Position und Orientierung bezüglich den ausgewählten starren Substrukturmitteln besitzt und die ausgewählten starren Substrukturmittel solche von den starren Substrukturmitteln sind.

Hierzu 12 Seite(n) Zeichnungen

- Leerseite -

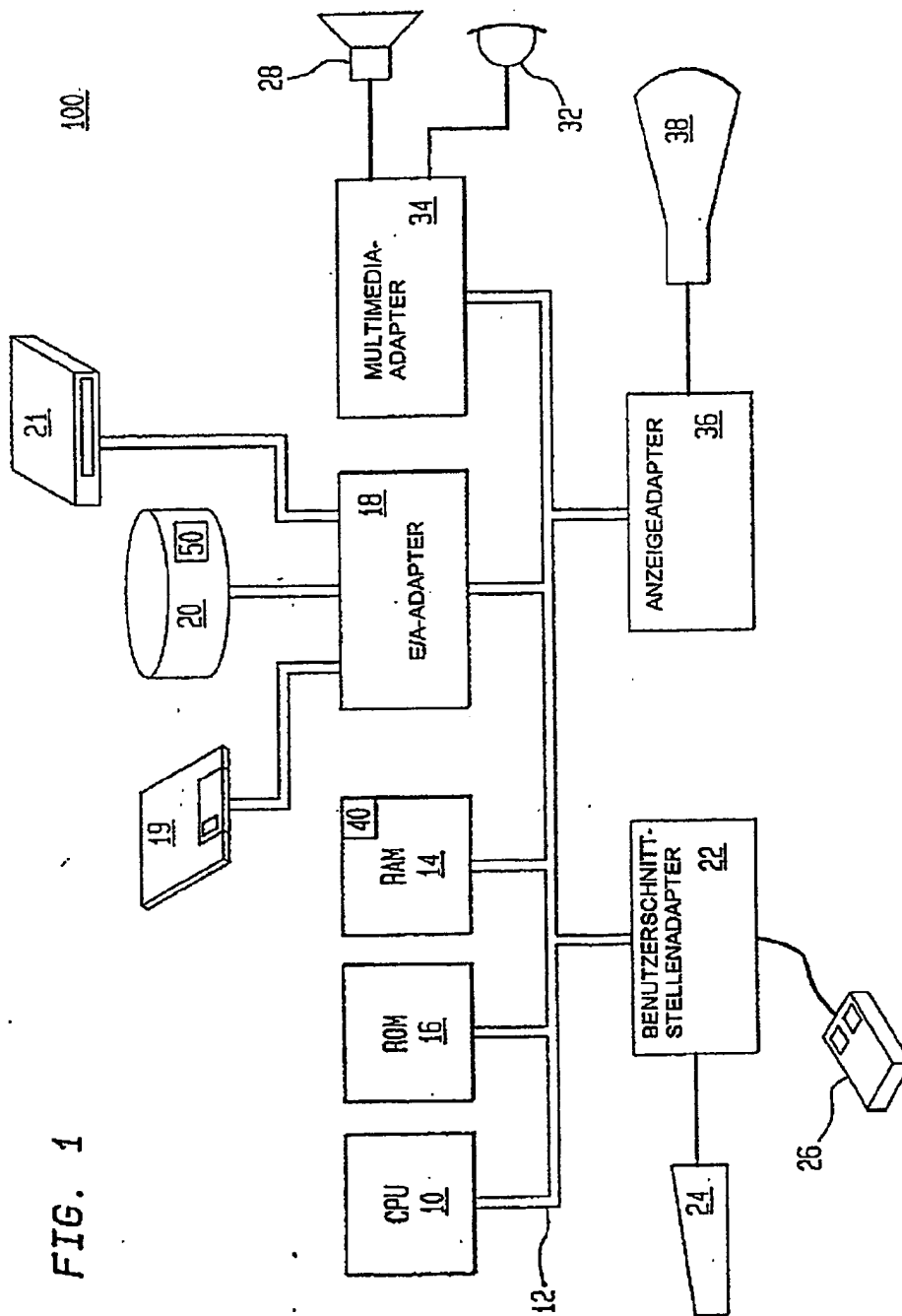


FIG. 1

FIG. 2B

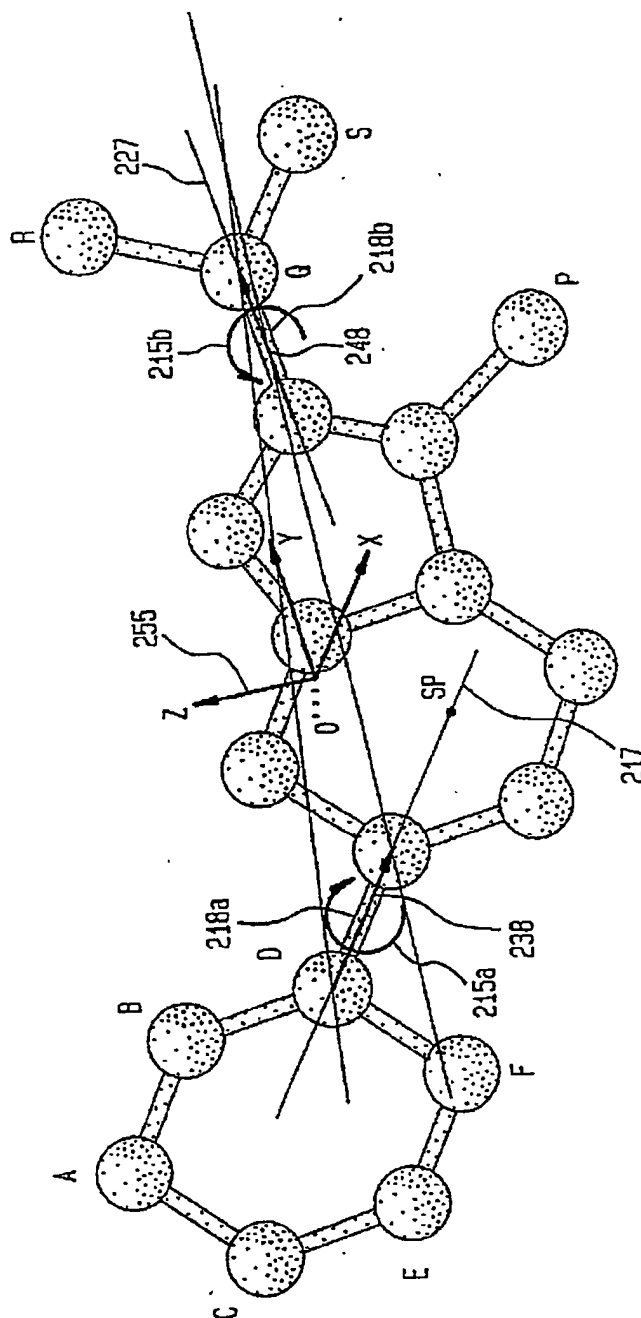


FIG. 2C

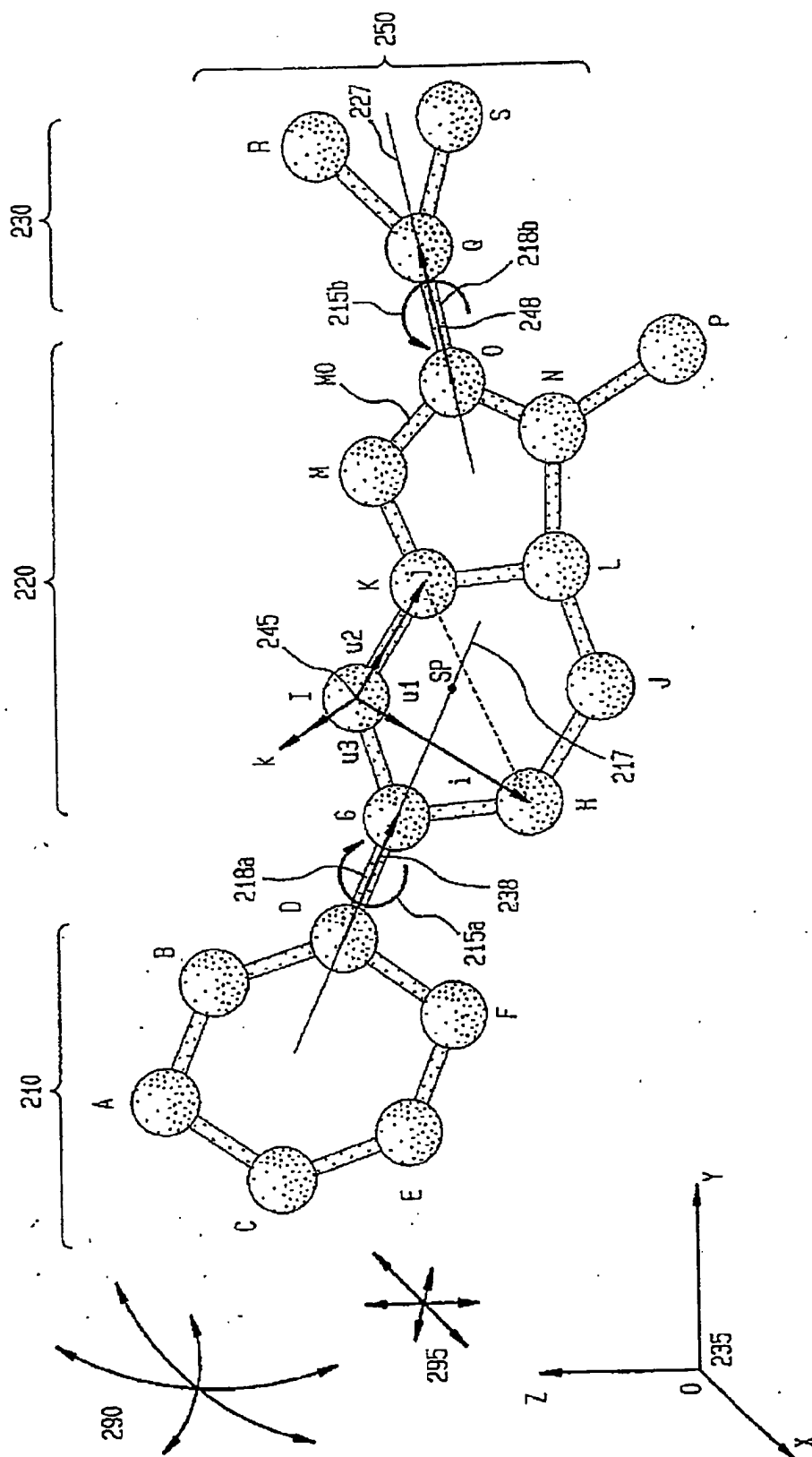


FIG. 3A

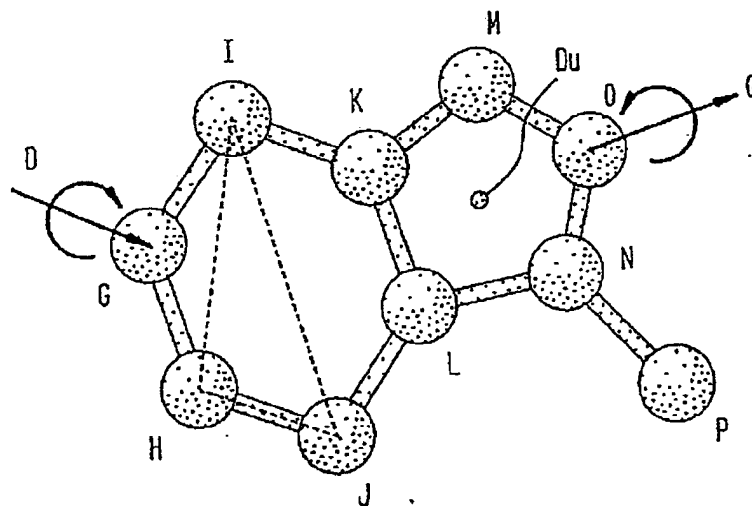
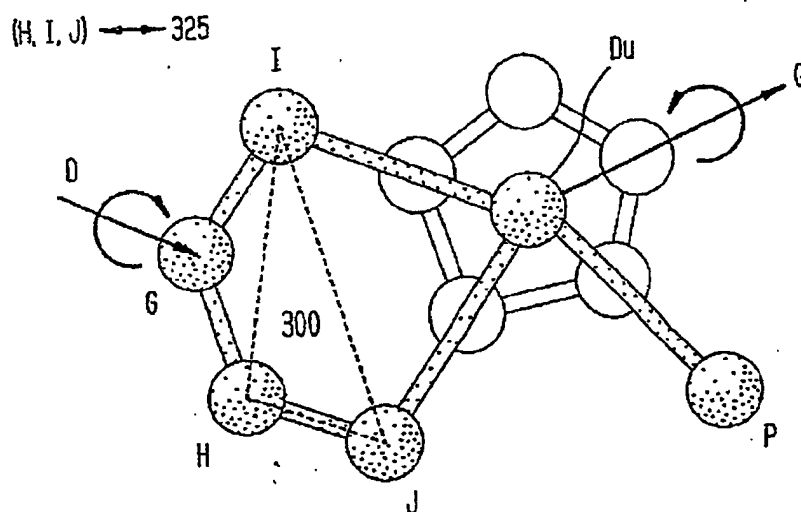


FIG. 3B



(H, I, J) \longleftrightarrow 335

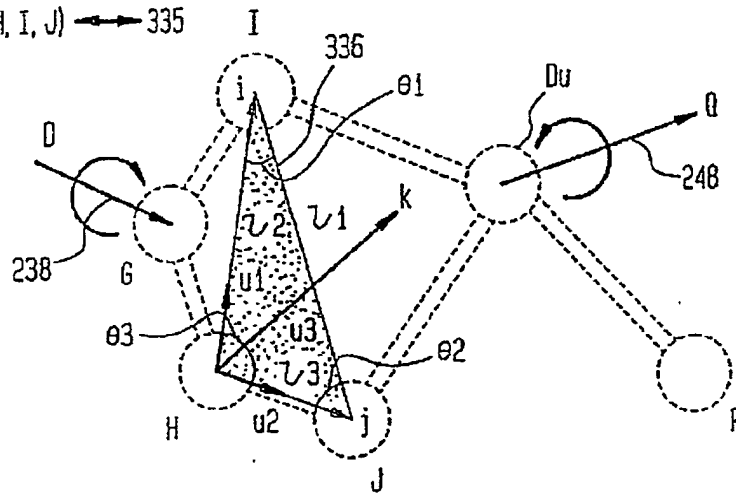


FIG. 3C

(I, J, Du) \longleftrightarrow 345

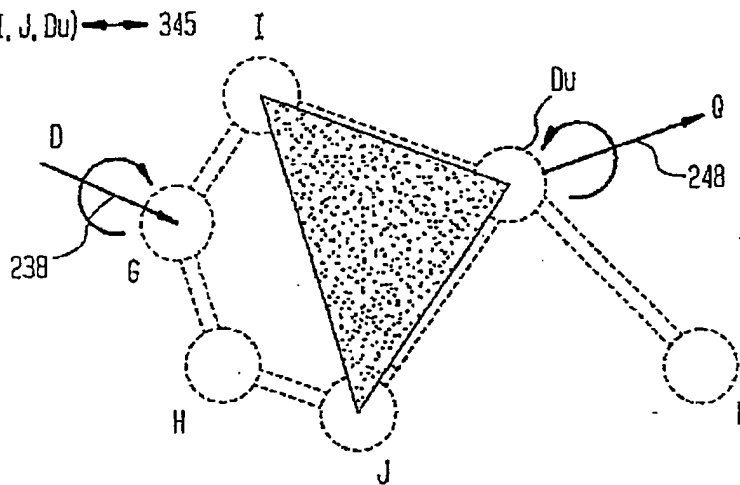


FIG. 3D

(D, G, I) \longleftrightarrow 355

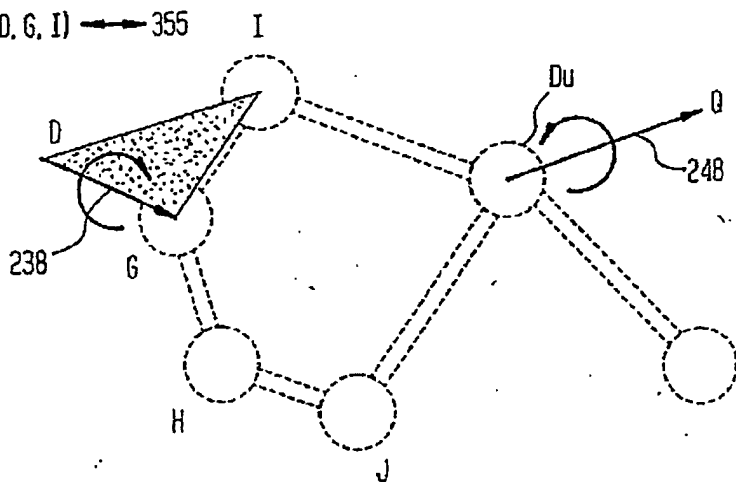


FIG. 3E

FIG. 4

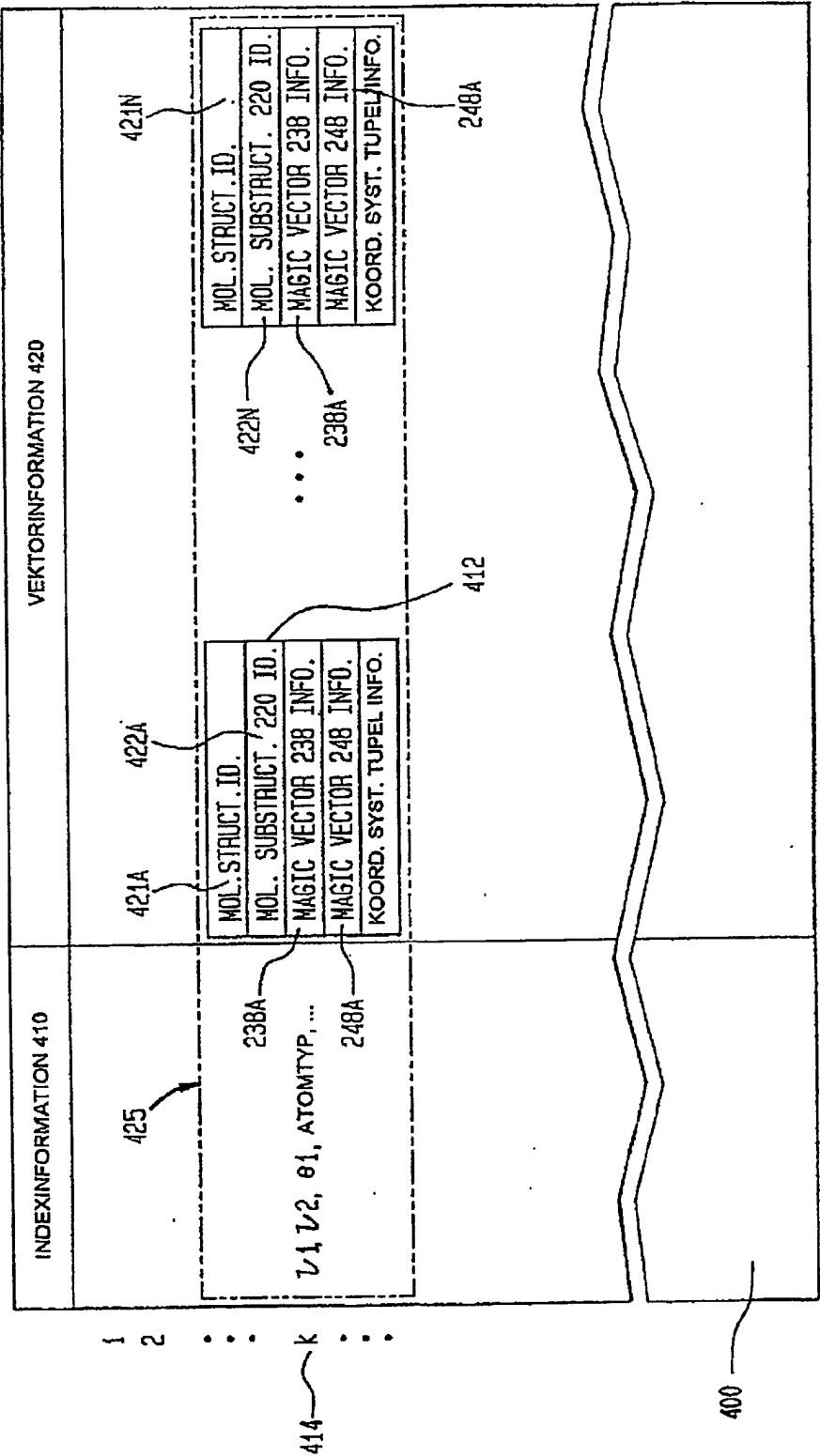


FIG. 5A

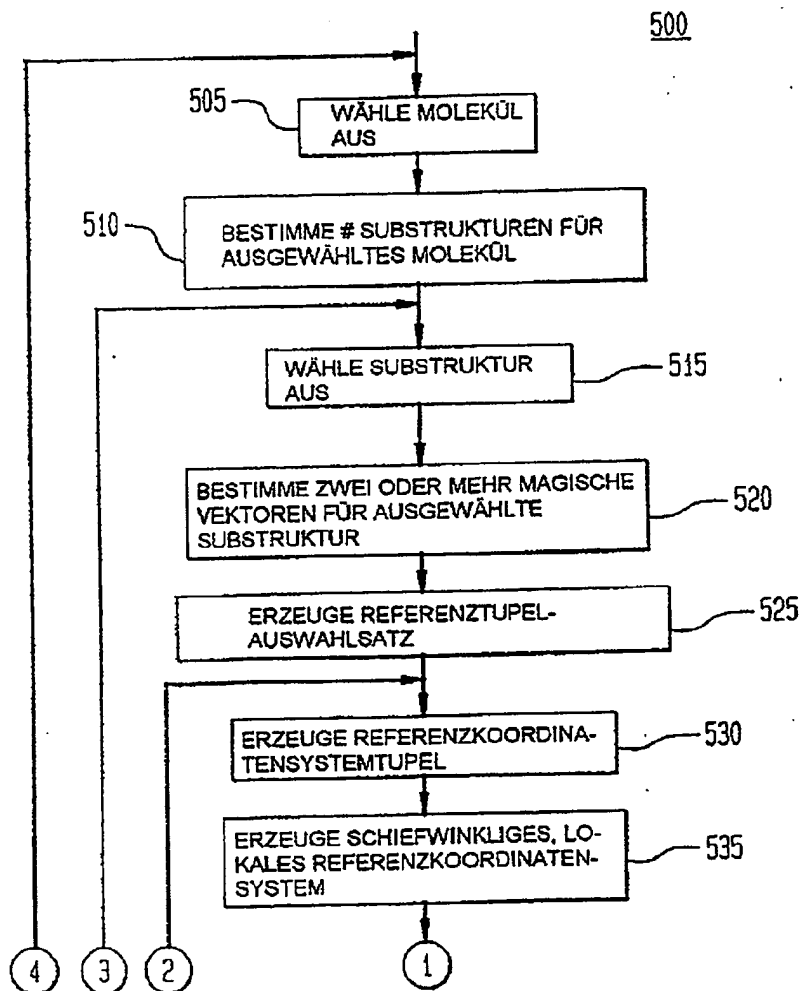


FIG. 5B

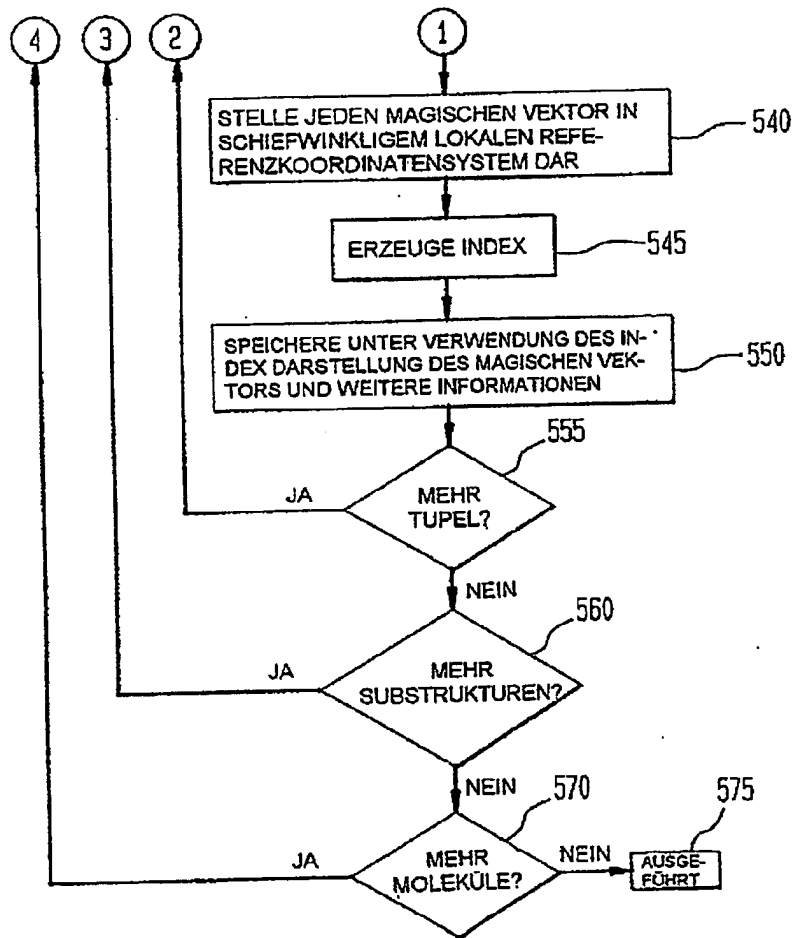


FIG. 6A

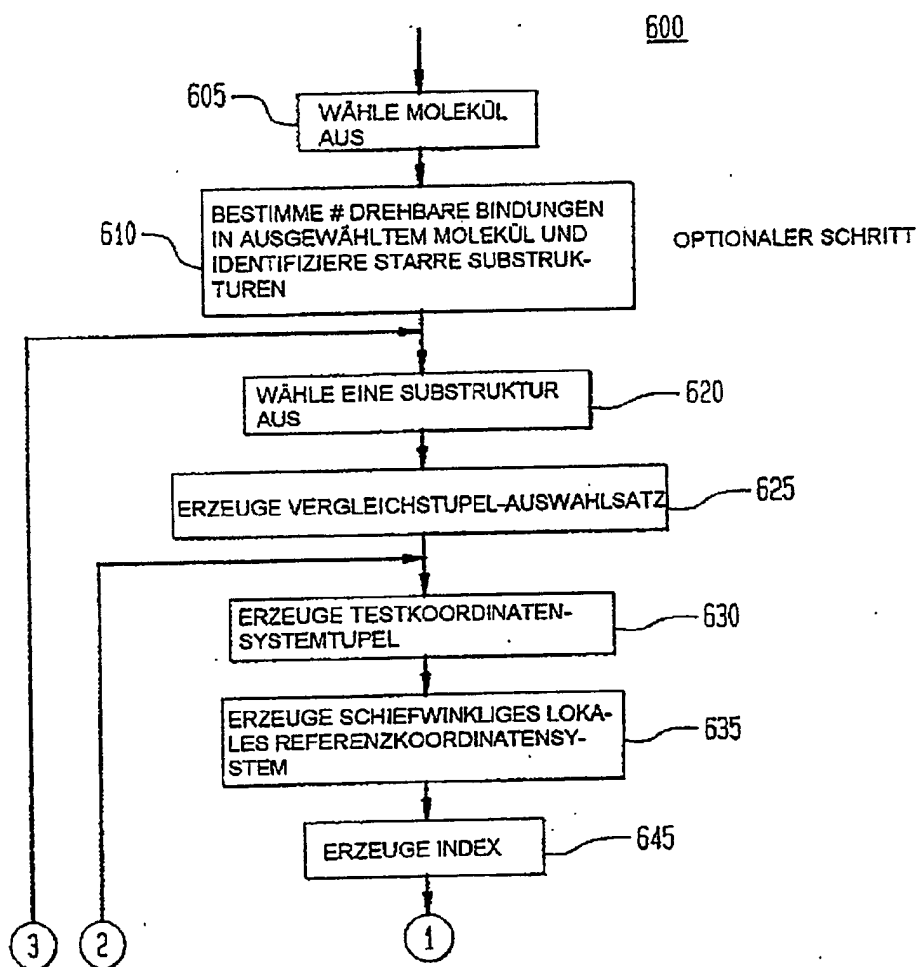


FIG. 6B

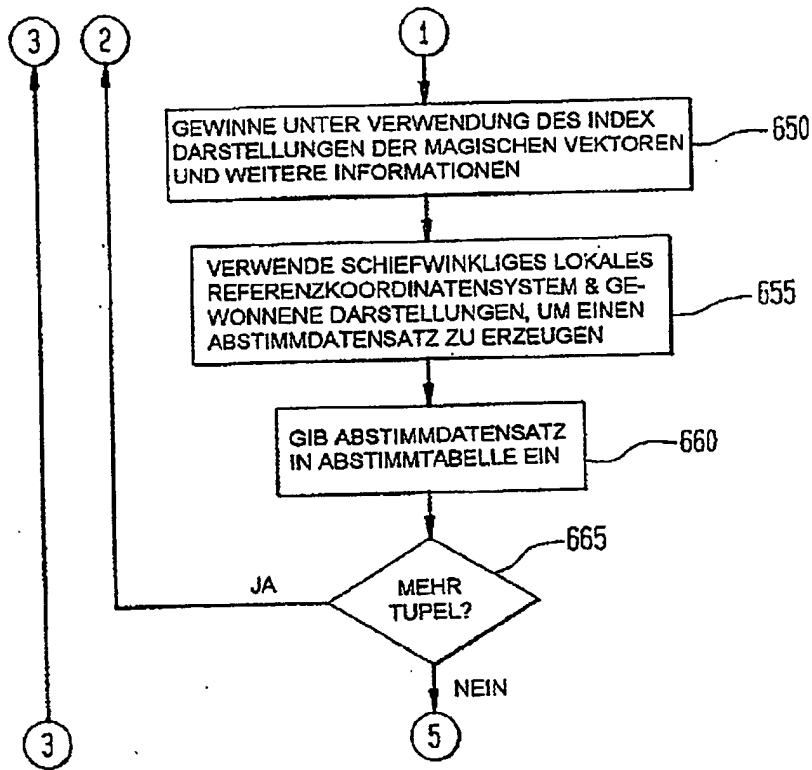


FIG. 6C

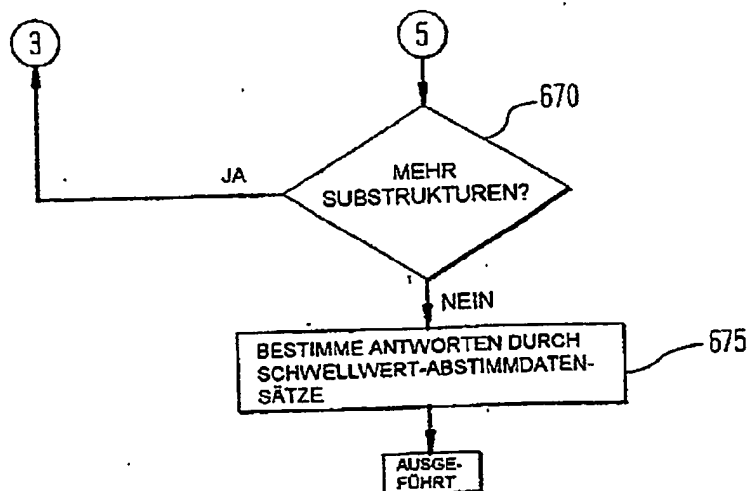


FIG. 7

